FACULDADE DE ECONOMIA DA UNIVERSIDADE DO PORTO

DISSERTAÇÃO DE MESTRADO

MODELAÇÃO, ANÁLISE DE DADOS E SISTEMAS DE APOIO À DECISÃO

Análise de Redes Sociais Aplicada a uma empresa de Retalho

Discente :

João Alexandre de Freitas Luís

Docente:

Prof. Dr. João Manuel Portela da Gama

Setembro 2015

Agradecimentos

Este trabalho não teria sido possível sem o apoio incondicional e extrema compreensão do meu orientador, Prof. João Gama, a quem muito agradeço.

Um palavra de agradecimento também, por todo o apoio que me disponibilizaram, aos meus pais e à Márcia, e aos meus filhos, pelo tempo que lhes foi subtraído.

Este trabalho foi apoiado pela Comissão Europeia no âmbito do projeto MAESTRA (Grant number ICT-2013-612944), a quem cabe também o meu agradecimento.

Índice

Índice de Tabelas

Índice de Ilustrações

Índice de Equações

1 Introdução

1.1 Motivação

O aparecimento da distribuição moderna favoreceu a recolha de volumes elevados de dados transacionais, permitindo a realização sustentada por informação rápida e relevante de análises de apoio a decisões comerciais, operacionais e de marketing, como são por exemplo, a localização de um produto na loja, o espaço ocupado na prateleira, o seu preço de venda, a operação logística de suporte ou as promoções realizadas localmente ou de forma centralizada.

Inicialmente, este tipo de dados, pela sua natureza transacional, permitiam relacionar a compra de um artigo com outro, desde que comprados na mesma transação. A análise de cabazes de compra (*Market Basket Analysis* - MBA) era, neste paradigma, a mais utilizada. Esta abordagem resultaria normalmente, por um lado, num volume excessivo de regras, por vezes redundantes, dificultando a sua análise, e por outro, numa influência exagerada dos artigos de elevada rotação.

Com o lançamento de cartões de fidelização, a descoberta de relações entre artigos deixou de estar dependente da transação. O vínculo pode agora ser baseado no cliente comum de dois ou mais artigos, contextualizado por um período de tempo. A análise de afinidade é, deste novo paradigma, o seu melhor exemplo, possibilitando também ações de marketing personalizadas, como, por exemplo, o envio de cupões customizados para o cliente.

Por outro lado, o elevado volume de dados que se torna necessário processar, exige a aplicação de estratégias analíticas que ajudem a uma produção de informação mais célere e eficaz.

Neste contexto, pretendemos demonstrar como a análise de redes sociais pode contribuir significativamente para a obtenção de informação sobre relações entre artigos e clientes, de forma mais concisa, contextualizada e mais interpretável. Mais especificamente, iremos incidir nos artigos de baixa rotação, habitualmente negligenciados pelas técnicas mais utilizadas.

1.2 Definição do Problema

A representação das relações entre os artigos através de uma rede facilita, como iremos ver, a deteção de grupos fortemente conectados e uma informação de mais fácil análise.

Um dos primeiros problemas que se levantam é o de estabelecer os fundamentos para identificar relações entre artigos. Trilhando a literatura existente, o seu grau de incidência conjunta, quer em transações como em clientes, foi o caminho que seguimos.

O segundo problema colocado foi o conjunto de pressupostos a utilizar para a construção da rede de artigos, tendo em vista que os métodos encontrados na literatura existente nos encaminhavam para uma excessiva concentração em torno dos artigos de

elevada rotação. O foco nos artigos de reduzida rotação, motivou-nos a testar estratégias de otimização com vista a obter redes menos densas e o mais abrangentes possível em termos de "cauda longa".

De ordem mais prática, deparámos com a necessidade de conhecer e adequar a base de dados utilizada para este trabalho, nomeadamente ao nível da identificação do tipo das lojas que a compunham, da resolução do problema da sazonalidade e do seu agrupamento em conjuntos tendencialmente homogéneos.

Por fim, colocámos como objetivo, desenvolver e aplicar fundamentos para a construção de redes de clientes. Neste propósito, levantam-se questões ao nível de como estabelecer um perfil de cliente, como estabelecer uma relação entre clientes, como agrupá-los e caracterizar esses agrupamentos.

1.3 Estrutura da Dissertação

O Capítulo 2 foca-se nos fundamentos teóricos da análise de redes sociais, elencando as métricas tradicionalmente utilizadas para a caraterização das redes, a nível de estrutura e de atores. O conceito de comunidade é abordado de igual forma neste capítulo, bem como a introdução ao tema da "cauda longa".

No capítulo 3, é dado relevo à literatura existente sobre redes de artigos, incluindo os métodos de obtenção destas redes, e em particular no artigo com base no qual partimos na construção de nova metodologia.

Os passos seguidos para o pré-processamento da base de dados, com vista à adequação ao trabalho a desenvolver, constituem a prólogo do capítulo 4, seguido pelo estudo do impacto dos artigos de elevada rotação na estrutura das redes de artigo. Este capítulo apresenta depois os resultados do trabalho realizado a três níveis:

- ao nível das relações entre artigos com base em transações comuns

- ao nível das relações entre artigos com base em clientes comuns

- ao nível das relações entre clientes com base em perfis de compra comuns

Por último, é incluído no capítulo 5 um resumo das dificuldades encontradas, das conclusões retiradas e dos caminhos possíveis de desenvolvimento deste trabalho.

2 Análise de Redes Sociais

A análise de redes sociais é uma área de pesquisa metodológica interdisciplinar, que recebeu contribuições de diversos ramos da Ciência (Gama et al, 2012), que teve origem na terceira década do século XX, como meio de conceptualização da estrutura das relações sociais estabelecidas entre um pequeno grupo de indivíduos.

Uma rede social consiste num número finito de atores (nós) e das relações que se estabelecem entre eles. São normalmente representadas por grafos (G), composto por vértices (V) e arestas (E).

A ordem de uma rede (n) é dada pelo número de vértices que a compõe (|V(G)|), e a sua dimensão (m) traduz-se no número de arestas nela contido (|E(G)|).

Podemos distinguir entre grafos direcionados ou não direcionados, consoante as ligações entre os nós possuem direção ou não. Num grafo não direcionado, as arestas apenas ligam arbitrariamente um nó ao outro, ao passo que num grafo direcionado podem existir entre dois nós ligações de sentido contrário.

Estas ligações podem ser ponderadas (ou pesadas) ou não, dependendo de existir ou não uma medida de força dessa relação. A existir esse grau de força, chamamos ao grafo de pesado.

Uma rede pode ser representada por uma matriz (nxn), em que cada interseção apresenta um valor que identifica, não só a existência de ligação, como o seu peso.

Em seguida são apresentadas as métricas que serão calculadas para as redes encontradas, nos casos que tal se justifique.

2.1 Métricas ao nível do nó

2.1.1 Grau ou Valência

Permite medir a envolvência de um nó na rede, através das suas relações primárias com outros nós, consistindo no número de nós que estão conectados com o nó analisado. Em redes direcionadas, podemos distinguir o suporte, como o número de nós com ligação a terminar no nó analisado, e a influência, como o número de nós com ligação a iniciar no nó analisado.

Esperamos encontrar nas redes de artigo uma distribuição de graus seguindo uma lei de potência, o que determinará a opção de focarmos a análise na cauda longa da distribuição, como veremos a seguir.

2.1.2 Centralidade de Intermediação

Indica o grau de importância de um nó na intermediação entre os outros nós da rede. Um valor elevado indica que se trata de um elemento vital na ligação entre diferentes regiões da rede.

2.1.3 Centralidade de Proximidade

Pretende medir a posição global de um ator na rede, através comprimento médio de todos os caminhos mais curtos entre um nó e os restantes nós da rede.

2.1.4 Centralidade do Vetor Próprio

Define-se como centralidade, não só a do próprio ator, mas a combinação linear das centralidades dos outros atores que a ele estão ligados.

2.1.5 Coeficiente de Agrupamento Local

Mede a coesão entre os vizinhos de um nó, através de um coeficiente de agrupamento. A transitividade presente numa rede leva a que a probabilidade de dois atores ligados a um nó estarem também ligados entre si seja mais elevada do que quando não existe um elo em comum.

2.2 Métricas ao nível da rede

2.2.1 Diâmetro e raio

O diâmetro representa a distância máxima existente entre qualquer par de nós da rede, ao passo que o raio a menor excentricidade encontrada na rede. Excentricidade de nó traduz-se no maior caminho mais curto (distância geodésica) que esse nó tem.

2.2.2 Distância Geodésica Média

Representa a média dos caminhos mais curtos existentes para todas as combinações de nós presentes na rede.

2.2.3 Grau Médio

Média de todos os graus de todos os nós da rede.

2.2.4 Reciprocidade

Mede a probabilidade de existência de simetria nas ligações estabelecidas entre pares de vértices.

2.2.5 Densidade

Quantifica o nível de conectividade presente na rede. Pode ser definida como a proporção de arestas em relação ao maior número de arestas possível.

2.2.6 Coeficiente de Agrupamento Global

É obtido pelo cálculo da média de todos os coeficientes de agrupamentos locais encontrados. Níveis elevados indicam habitualmente uma maior probabilidade de aparecimento de cliques (conjunto de nós com ligações entre todos os seus elementos).

2.3 Deteção de Comunidades

A distribuição das arestas ao longo de uma rede é muito heterogénea, o que leva ao aparecimento de zonas mais densas separadas por zonas menos densas. Dizemos por isso que existe uma estrutura de comunidade.

Existem algumas metodologias para a deteção de comunidade em redes sociais (Newman et al, 2004), mas no âmbito deste trabalho apenas iremos utilizar aquela baseada na otimização da modularidade.

A modularidade de uma rede (Q) é definida por:

$$Q = \sum_i (e_{ii} - a_i^2)$$

Equação 1 - Equação para o cálculo da modularidade de uma rede

onde e_{ii} é a proporção de arestas que ligam dois nós da mesma comunidade i , e a_i representa a proporção de términos de ligações que se encontram em nós da comunidade *i*.

Quanto maior a modularidade, mais eficaz é a divisão da rede em comunidades, pelo que se trata do objetivo a otimizar. Este processo de otimização é muito exigente em termos de processamento, recorrendo-se a heurísticas que reduzem o tempo de otimização (Fortunato, 2010), como é o caso do *software* Gephi.

2.4 "The Long Tail"

Uma das tendências mais marcantes do retalho nos dias de hoje é a progressivo aumento do peso na vendas dos artigos de menor rotação. A regra 80/20, que estabelece o pressuposto de que 20% dos artigos mais vendidos representam 80% das vendas, tem perdido a sua relevância devido sobretudo à diminuição do peso nas vendas representado pelos artigos de alta rotação

O aparecimento de grandes empresas de retalho *online*, como a Amazon ou Rhapsody, com gamas de artigos mais extensas que o comércio offline, possibilitadas pela desmaterialização ou centralização de stocks e pela libertação das restrições que os mercados locais representam para a definição de gama (p.e. uma loja alimentar situada numa pequena comunidade terá de gerar tráfego baseando-se em produtos de grande consumo, visto estar vocacionada para um mercado reduzido), é um dos principais fatores que contribuem para este fenómeno (Anderson, 2006).

Por outro lado, a tecnologia que permite a aproximação entre pessoas com interesses comuns, embora afastados do *mainstream*, favorece o aparecimento e fortalecimento de

sub-culturas, transversais geograficamente, que por sua vez, representam mais oportunidades de aparecimento de artigos vocacionados para nichos de mercado.

Podemos falar de uma democratização dos meios de procura e oferta, e de uma comunicação entre eles facilitada pelos meios tecnológicos (Anderson, 2006), que resulta numa distribuição de vendas de cauda longa. Ou seja, a "regra dos 98%", em que apenas 2% dos artigos apresentam uma frequência muito alta em oposição à frequência baixa dos restantes artigos.

Sempre que possível iremos realizar as análises à luz desta nova realidade, tendo em consideração que o esforço analítico do negócio do retalho era no passado focalizado nos artigos de frequência superior, de forma a rentabilizar o esforço de vendas. Essa estratégia tem ignorado o potencial da cauda longa, o que tentaremos evidenciar neste estudo.

3 Redes de Artigos

3.1 *Market Basket Analysis* (MBA)

Uma das ferramentas mais importantes da MBA - regras de associação - permite evidenciar relações de compra entre artigos "escondidas" na base transacional de um retalhista, que possam mais tarde ser "acionadas" ao nível comercial e de marketing.

Em resumo, trata-se de conhecer qual a propensão a encontrar no mesmo cabaz de compras um produto B (item consequente), sabendo que contém o produto A (item antecedente). A mesma lógica poderá ser aplicada a grupos de artigos (*itemsets*).

A aplicação de regras de associação resulta normalmente num elevado número de regras, muitas vezes irrelevantes ou redundantes, submetidas a triagem baseada num limite inferior para o seu suporte (avaliando o seu poder de generalização) e para seu nível de confiança (medindo a força da regra). Após esta triagem, assiste-se ainda assim a um volume elevado de regras que torna difícil a identificação de relações interessantes (Chawla at al, 2011). Em grande medida, as medidas de interesse das regras dependem do conhecimento do negócio.

O objetivo da aplicação da SNA (Social Network Analysis) a este tipo de análise é o de pretender facilitar a deteção de comunidades de artigos com forte relação entre si, ultrapassando as limitações das regras de associação (Chawla et al, 2011).

Na representação de produtos em rede, a ligação entre os nós (artigos) estabelece-se com base no número de vezes em que os dois artigos aparecem na mesma transação. Uma ligação baseada apenas numa incidência em comum, para além de originar redes de produtos demasiado densas, pode abarcar relações entre artigos muito pouco representativas. Com vista a expurgar incidentes esporádicos, estabeleceremos um número mínimo de compras em comum para existir uma ligação entre dois artigos.

A força da ligação irá depender da força de relação, ou seja, do poder preditivo que o antecedente tem face ao consequente, medido através do nível de confiança. Traduz-se pelo quociente da probabilidade de ocorrência simultânea dos dois artigos numa transação pela probabilidade de ocorrência do artigo antecedente (ver tabela 1). Esta métrica é por definição assimétrica, e originará ligações nos dois sentidos de força diferente.

Transações Totais	t
Frequência artigo A	a
Frequência artigo B	b
Frequência conjunta A e B (Suporte)	c
Suporte Relativo	c/t
Confiança (A -> B)	c/a
Confiança (B -> A)	c/b

Tabela 1 - Métricas MBA

Estudos prévios evidenciaram que a presença de artigos com grande frequência, que têm ligações com a quase totalidade dos restantes, torna a rede muito densa e difícil de analisar.

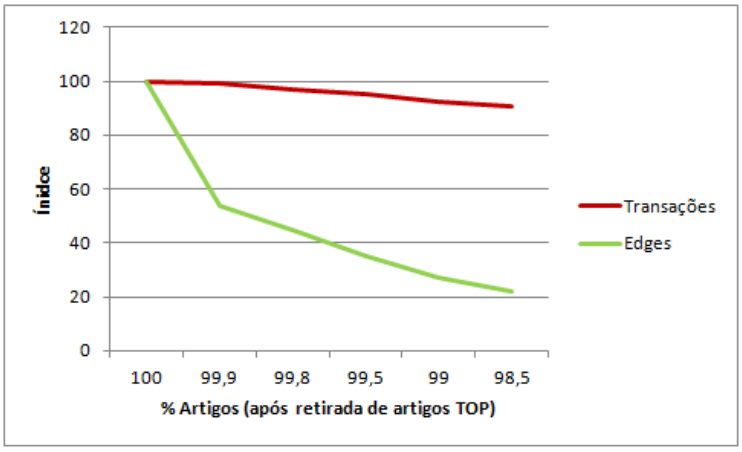

Ilustração 1 - Sensibilidade de uma rede de artigos à presença de artigos de elevada frequência

Podemos ver na ilustração 1, que à medida que retiramos os artigos de maior frequência na construção de uma rede de artigos, para além do efeito nas transações consideradas, obtemos uma rápida diminuição do número de ligações entre artigos, logo redes menos densas e mais fáceis de analisar. No caso representado, apenas retirando os 1,5% artigos mais frequentes reduzimos o número de ligações em cerca de 80%.

3.2 Comunidades de Artigos

Em SNA, uma comunidade representa um grupo de nós em que as ligações entre si são mais fortes que as relações com os membros das outras comunidades.

A aplicação do conceito de comunidades a uma rede de artigos tem em vista isolar sub-redes de artigos com ligações fortes entre si. Pretende-se assim obter uma decomposição do universo geral dos artigos em sub-grupos (ver ilustração 2) que apresentam uma forte interdependência, eliminando assim o problema de redundância que a utilização das regras de associação muitas vezes apresenta.

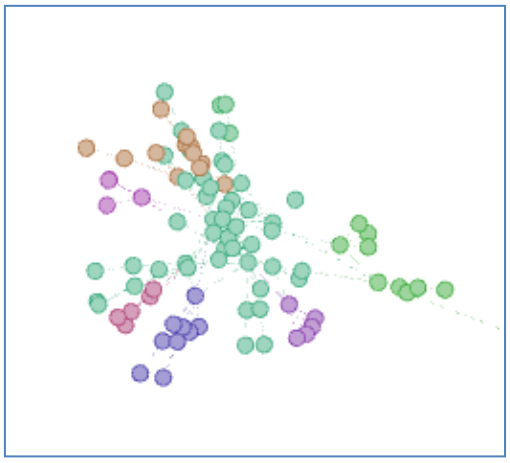

Ilustração 2 - Detalhe de um conjunto de comunidades encontradas numa rede de artigos

Para a construção de comunidades existem dois parâmetros de especial relevância e que irão ter impacto na estrutura de comunidades a encontrar na rede:

- o nível de suporte mínimo (irá variar o nº de ligações existentes entre os nós);

- o grau de confiança da relação (maior ou menor poder preditivo).

O método a utilizar para a construção das redes de artigos (Chawla et al, 2011) irá apoiar-se no uso de um nível de suporte mínimo inicial baixo, incrementando esse parâmetro por etapas e avaliando a evolução da modularidade. O nível de modularidade deverá atingir um valor estável a partir do qual novos incrementos no parâmetro não terão efeito significativo, podendo inclusivamente regredir (ver ilustração 3). Para diferentes valores do parâmetro que apresentem modularidade máxima, será escolhido o que permita uma maior modularidade, pois será aquele que possibilita a manutenção de um maior número de artigos e ligações.

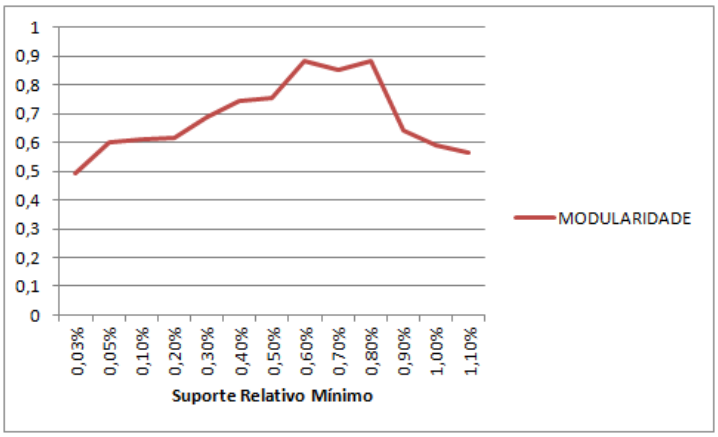

Ilustração 3 - Sensibilidade da Modularidade a variações no Suporte Mínimo

Numa segunda fase, procederemos à avaliação da informação contida em cada comunidade.

3.3 Medição da Utilidade das Comunidades

Seguindo a metodologia proposta por Chawla (Chawla et al., 2011) para a medição do valor informativo de uma comunidade de artigos, uma comunidade útil deve ser suficientemente grande para oferecer informação sobre o comportamento de compra do cliente, mas suficientemente pequena para ser interpretável.

O valor de informação de uma comunidade centra-se no somatório dos níveis de confiança de todas as ligações existentes (Ei) entre os nós da comunidade (Gi):

$$I(G_i) = \sum_{(p1,p2)\in E_i} P(p1|p2)$$

Equação 2 - Equação para o cálculo da informação obtida numa comunidade de artigos

Existe uma grande diversidade de medidas que poderiam ser utilizadas para representar a força da ligação (Pang-Ning Tal et al, 2004), mas o nível de confiança tem como vantagem poder variar apenas entre 0 e 1, evitando-se assim valores extremos.

Desta forma, uma comunidade com um maior número de nós, terá tendencialmente mais ligações, e terá um valor de informação maior do que uma comunidade pequena. De forma a relativizar este valor pelo número de nós da comunidade (Vi), construiremos a métrica densidade de informação D(Gi), espelhada na seguinte fórmula:

$$D(G_i) = \frac{I(G_i)}{|V_i|}$$

Equação 3 - Equação para o cálculo da densidade de informação obtida numa comunidade de artigos

A utilidade total da comunidade será determinada então pela média harmónica das métricas atrás determinadas:

$$U(G_i) = \frac{2 \times I(G_i) \times D(G_i)}{I(G_i) + D(G_i)}$$

Equação 4 - Equação para o cálculo da utilidade de uma comunidade de artigos

3.4 Redes de Afinidade (Redes de Produtos baseada em clientes comuns)

As redes de artigos baseadas em clientes comuns são construídas a partir dos clientes partilhados entre dois artigos durante um período de tempo (Hyea Kyeong Kim, 2012). A força da ligação depende novamente da capacidade preditiva do antecedente sobre o consequente.

A densidade das redes baseadas em clientes comuns é normalmente superior do que no caso das redes baseadas em transações comuns. Uma das características destas é a de que os artigos menos partilhados têm uma maior probabilidade de estarem ligados a outros artigos, visto não dependerem de estar presentes na mesma transação, mas no conjunto de compras do período do mesmo cliente, favorecendo assim o aparecimento de "cauda longa".

Uma das diferenças entre os dois tipos de redes encontradas em estudos anteriores é a de que os artigos mais importantes em termos de centralidade nas redes baseadas em MBA representam os artigos mais frequentes, de compra diária, ao passo que nas redes de afinidade (clientes comuns) representam os artigos mais comuns a todas as famílias.

3.5 Contexto Multi-Lojas vs Loja / Sazonalidade

Quando se procuram encontrar padrões de compra num período alargado de tempo e num conjunto de lojas enfrentamos dois tipos de problemas (Yean-Liang Chen et al, 2005) :

- os produtos sazonais e promocionais que apenas estão presentes nas lojas durante um período limitado de tempo

- gamas diferentes entre lojas implicam que cada par de produtos considerados seja contextualizado de formas distintas

Para contornar o primeiro problema, optámos por incluir apenas artigos disponíveis em todos os 12 meses. Consideramos que um artigo esteve disponível na loja i e no mês j se tivermos pelo menos uma venda nesse mês e nessa loja.

Quanto ao segundo, gamas diferentes entre lojas, utilizámos para efeito de cálculo do suporte relativo, todas as transações das lojas em que os dois artigos que compõe cada par estiveram presentes (medido novamente pela existência de vendas). Desta forma, obtivemos diferentes contextos em termos de transações (ver tabela 2), adequando-se a cada caso, e não um contexto universal de transações.

		Lojas				
		A	B	C	D	Total
Transações Totais		1500	500	2000	2500	6500
Frequência Absoluta	Artigo 1	200	50	300	400	950
	Artigo 2	100	30	150		280
	Artigo 3	30		20	30	80
Suporte Absoluto	Artigos 1 e 2	50	5	50		105
	Artigos 2 e 3			10		10
	Artigos 1 e 3	10		5		15
Suporte Relativo sem considerar o contexto	Artigos 1 e 2	105 / 6500 = 1,62%				
	Artigos 2 e 3	10 / 6500 = 0,16%				
	Artigos 1 e 3	15 / 6500 = 0,23%				
Suporte Relativo contextualizado	Artigos 1 e 2	105 / (1500+500+2000) = 2,63%				
	Artigos 2 e 3	10 / (1500+2000) = 0,29%				
	Artigos 1 e 3	15 / (1500+2000+2500) = 0,25%				

Tabela 2 - Exemplo do ajuste no cálculo do suporte relativo provocado pela contextualização multi-loja

3.6 Redes de Clientes

As redes de clientes baseiam-se na proximidade existente entre os perfis de compras de dois clientes (estilos de vida) (V. Miguéis, 2012), medida através duma métrica que represente a distância entre os dois, em termos de atributos relevantes.

Uma questão importante é a da escolha dos atributos que são relevantes para esta função. Para esta tarefa utilizaremos uma metodologia baseada na extração das categorias que mais informações fornecem entre a separação de clientes diferentes, com base nas suas compras. Esse objetivo é atingido através da aplicação da análise de componentes principais aplicadas às categorias de artigos.

Para vários níveis de suporte (medido em *overlap* de escolhas nos atributos relevantes), iremos detetar comunidades e analisar o grau de modularidade da rede, à semelhança do efetuado antes para os artigos.

4 Estudo de Caso e Análise de Resultados

4.1 Introdução

Todas as análises efetuadas neste estudo serão baseadas numa base de dados transacionais de uma empresa de retalho portuguesa, recolhidos por um período de um ano, num total de três dezenas de milhão de linhas de transação, representando cerca de dois milhões e meio de transações, realizadas por sessenta e quatro mil clientes (amostra de cerca de dois por cento do total de clientes), em cento e cinquenta mil artigos diferentes, e em cerca de quinhentas e cinquenta lojas.

4.2 Ferramentas Utilizadas

Devido ao elevado número de cruzamentos necessários para processar as frequências comuns, foi utilizado o software *Sql Server*.

Após um trabalho inicial de desenho dos cruzamentos necessários, numa primeira fase em apenas uma loja, foi necessário otimizar os *queries* através da utilização de índices por forma a tornarem mais rápido a sua execução.

As análises de sensibilidade implicaram a utilização do software R, mais versátil para a execução de ciclos, com ligação à base de dados em *Sql*.

Para a apresentação e cálculo de métricas de redes, recorremos ao *software* Gephi.

A análise em componentes principais de suporte à rede de clientes foi efetuada em SPSS.

4.3 Pré-processamento

4.3.1 Exclusão de aberturas e fecho de lojas

Das 550 lojas representadas no *dataset* optámos por excluir as lojas que não dispunham de dados referentes a todos os meses, evitando assim aberturas e fechos de lojas, ficando assim com um total de 434 lojas.

4.3.2 Grupos de lojas baseados em perfis de tipo de artigos vendidos

Após este passo, analisámos o perfil de cada loja em termos de produtos vendidos, com base no nível superior de estrutura mercadológica (neste caso a direção comercial - DC), considerando apenas as DC´s que representam 98% das vendas das lojas (com vista a eliminar gamas residuais).

Identificámos desta forma 6 grupos de lojas por perfil de venda DC, como podemos ver na tabela 3.

Grupo	# Lojas	Volume Vendas Grupo	Venda Média por Loja	Nº médio de transações por Loja	Transação Média no Grupo	DC´s principais
1	1	966 230	966 230	27 771	34,8	10,11,27
2	103	55 264 128	536 545	16 362	32,8	10,11
3	75	15 501 040	206 681	8 933	23,1	10
4	68	150 378	2 211	902	2,5	12
5	50	413 434	8 269	1 016	8,1	23
6	137	1 858 111	13 563	782	17,3	27

Tabela 3 - Grupos de Loja por Perfil DC

Dos 6 grupos encontrados, optámos por focar a análise nos grupos 2 e 3, porque apresentam um nº de transações por loja e transação média superior, favorecendo o aparecimento de um maior número de relações entre artigos, para além de se tratar de lojas de dimensão superior, logo com uma gama mais extensa, que à partida garante o aparecimento de "caudas longas".

4.3.3 Classificação com base em gamas comuns

Após a seleção dos grupos 2 e 3, totalizando 178 lojas, construímos para cada grupo uma rede de lojas cujas ligações são estabelecidas pela distância entre gamas de lojas diferentes.

A metodologia utilizada foi a de considerar a gama de uma loja como a representação de todos os artigos com pelo menos uma venda durante o ano, tratando-se de uma aproximação à gama real da loja (informação que estava indisponível). A distância entre gamas foi medida através do índice de Jacquard. Considerando G_a e G_b como a totalidade de artigos que compõe as gamas das lojas a e b, e G_{ab} como a interseção das duas gamas, o índice será calculado com base na fórmula:

$$G_{ab}/(G_a + G_b - G_{ab})$$

Equação 5 - Índice de Jacquard aplicado a gamas de artigos vendidos nas lojas

Como limite mínimo para considerar duas gamas como semelhantes, após algumas tentativas, chegámos ao valor de 0,2.

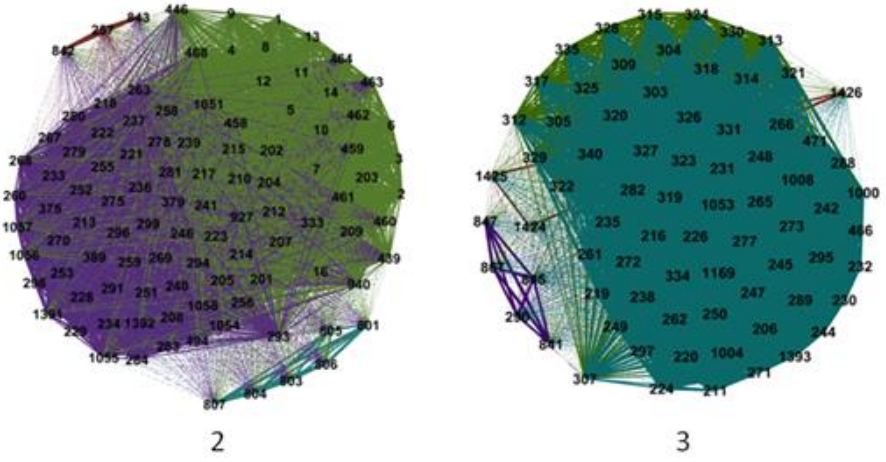

Ilustração 4 - Redes de Lojas com base em gamas em comum (Min 20%)

Para cada grupo foi construída uma rede de lojas (conforme ilustração 4) e aplicámos o algoritmo de deteção de comunidades presente no *software Gephi*. Os *clusters* resultantes estão identificados na tabela 4. Optámos por excluir os clusters com menor número de lojas, selecionando os seguintes:

- Cluster BIG (2_2) - lojas de maior dimensão, gama extensa, maior transação média.

- Clusters MEDHIGH (2_3) e MEDLOW (3_3) - Lojas de dimensão média, gama média, distinguem-se pelo tipo de produtos presentes em gama.

- Cluster SMALL (3_2) - Lojas de dimensão mais pequena, gama estreita, transação média baixa.

Serão estes quatro *clusters* a base das análises subsequentes.

Grupo	# Lojas	Volume Vendas Grupo	Venda Média por Loja	Nº médio de transações por Loja	Transação Média no Grupo	Nº Médio Artigos em Gama
2_0	6	2 590 578	431 763	14 090	30,6	14 334
2_1	3	1 077 788	359 263	12 045	29,8	12 926
2_2 BIG	36	34 302 290	952 841	26 564	35,9	27 475
2_3 MEDHIGH	58	17 293 472	298 163	10 487	28,4	13 189
3_0	3	22 526	7 509	690	10,9	435
3_1	5	1 387 622	277 524	9 731	28,5	10 205
3_2 SMALL	26	2 800 056	107 694	7 651	14,1	6 241
3_3 MEDLOW	41	11 290 836	275 386	10 253	26,9	11 854

Tabela 4 - Caracterização dos clusters encontrados através do algoritmo de modularidade

4.4 Caracterização dos *clusters*: métricas de rede e impacto de top vendas

A construção das redes de artigos e análise de impacto dos artigos top vendas foi baseada no período de vendas do primeiro bimestre. A opção por este período de vendas deveu-se a restrições de processamento. A construção de redes de artigos para um período anual considerando todos os artigos revelou-se muito exigente para os recursos de processamento disponíveis.

Assim, com base no primeiro bimestre, construímos as redes de produto para cada cluster e analisámos o impacto dos artigos top vendas. Para evitarmos redes demasiado grandes e relações entre artigos pouco representativas, optou-se por fixar um suporte mínimo relativo de 0,01% e absoluto de 10 transações.

Por outro lado, após alguns testes preparatórios, detetaram-se alguns casos de ligações que se baseavam em transações conjuntas de artigos efectuadas por um único cliente. Isto deve-se à existência de clientes com características "especiais", que incidem as suas compras num número reduzido de artigos, mas com compras de quantidades muito elevadas. Para contornar este problema, decidimos um limite mínimo de 2 clientes

distintos, ou seja, se todas as transações conjuntas entre o artigo A e B são de apenas um cliente, essa relação é filtrada da análise.

Podemos ver na tabela 5 algumas métricas de rede calculadas para as redes encontradas.

Cluster	NÓS	LIGAÇÕES	GRAU MÉDIO	DIÂMETRO	MODULARIDADE	DENSIDADE
BIG	6301	131374	41,699	7	0,581	0,007
MEDHIGH	3896	80027	41,082	7	0,479	0,011
SMALL	758	5266	13,894	6	0,627	0,018
MEDLOW	2832	55051	38,878	6	0,456	0,014

Tabela 5 - Métricas das redes de produto de cada cluster

4.5 Impacto dos artigos mais vendidos

Para medir o impacto dos artigos mais vendidos, analisámos em que medida a retirada da análise dos artigos top vendas afeta a rede de artigos e suas métricas.

Na tabela 6 podemos ver que mesmo apenas considerando os 100 top vendas, temos presenças entre 60 a 69% das transações, mas o seu impacto em linhas de transação é menor (de 18 a 25%).

	Transações com presenças		Linhas de transação		Ligações	
Cluster	TOP 100	TOP 200	TOP 100	TOP 200	TOP 100	TOP 200
BIG	60%	67%	18%	25%	76%	92%
MEDHIGH	67%	74%	21%	29%	82%	95%
SMALL	66%	75%	25%	35%	96%	98%
MEDLOW	69%	76%	23%	31%	86%	96%

Tabela 6 - Impacto dos artigos top vendas em transações e ligações de rede

Nas tabelas 7 e 8, podemos ver que obtivemos redes mais pequenas, menos densas, mas de diâmetro superior, favorecendo uma estrutura mais modular.

Cluster	NÓS	LIGAÇÕES	GRAU MÉDIO	DIÂMETRO	MODULARIDADE	DENSIDADE
BIG	3202	31882	19,914	11	0,716	0,006
MEDHIGH	1915	14641	15,291	12	0,669	0,008
SMALL	232	212	1,828	13	0,957	0,008
MEDLOW	1361	7710	11,33	9	0,678	0,008

Tabela 7 - Impacto da retirada dos artigos top 100

Cluster	NÓS	LIGAÇÕES	GRAU MÉDIO	DIÂMETRO	MODULARIDADE	DENSIDADE
BIG	2848	10913	7,664	13	0,868	0,003
MEDHIGH	1613	4003	4,963	12	0,876	0,003
SMALL	138	93	1,348	5	0,969	0,01
MEDLOW	1065	2028	3,808	18	0,901	0,004

Tabela 8 - Impacto da retirada dos artigos top 200

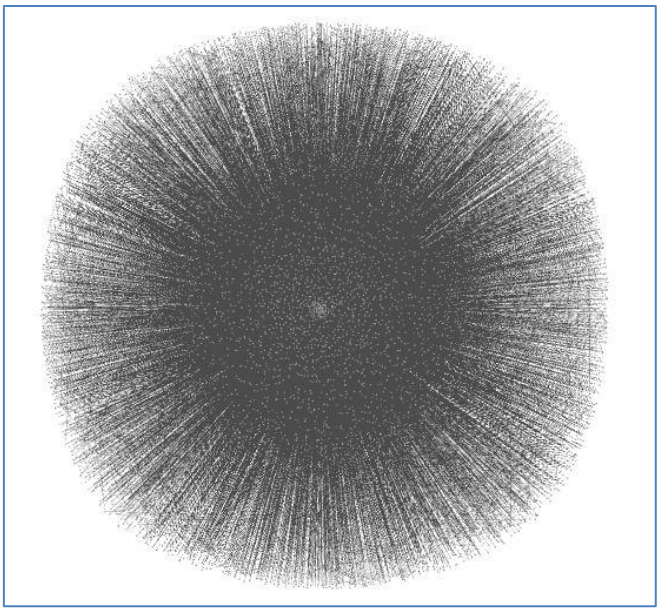

Ilustração 5 - Rede de artigos cluster BIG no bimestre 1

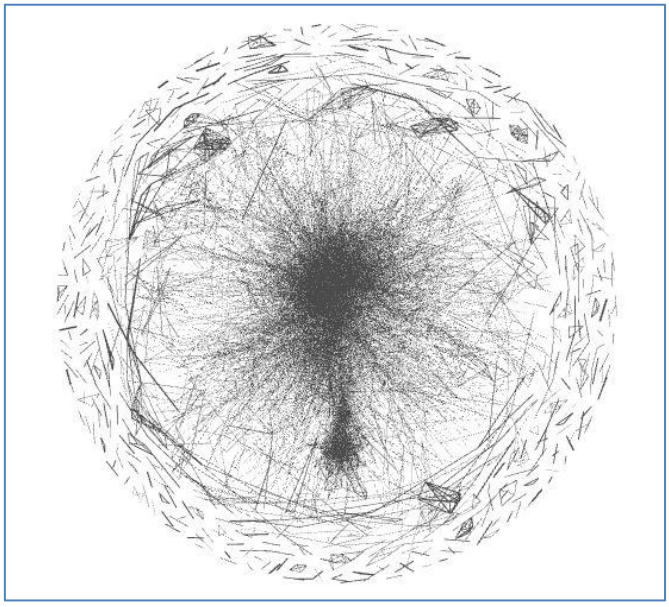

Ilustração 6 - Rede de artigos do cluster BIG no bimestre 1 após retirada top 200 vendas

No *cluster* BIG, a eliminação dos 200 artigos de maior frequência, resultou numa rede com menos de metade dos nós (6301 vs 2848: menos 55%), com apenas 8% das ligações (131374 vs 10913). Esta diminuição em termos de ligações favoreceu o aparecimento de uma estrutura em comunidades, resultando numa modularidade superior (0,581 vs 0,868). A eliminação dos artigos de elevada frequência provocou uma queda substancial no grau médio (41 vs 7,6).

Graficamente (ver ilustrações 5 e 6), de uma rede com um grande núcleo de artigos muito frequentes, a partir dos quais irradiavam ligações para a quase totalidade dos menos frequentes, passamos para uma rede com dois núcleos (um menor e um maior), mas constituída principalmente por pequenos componentes de poucos artigos.

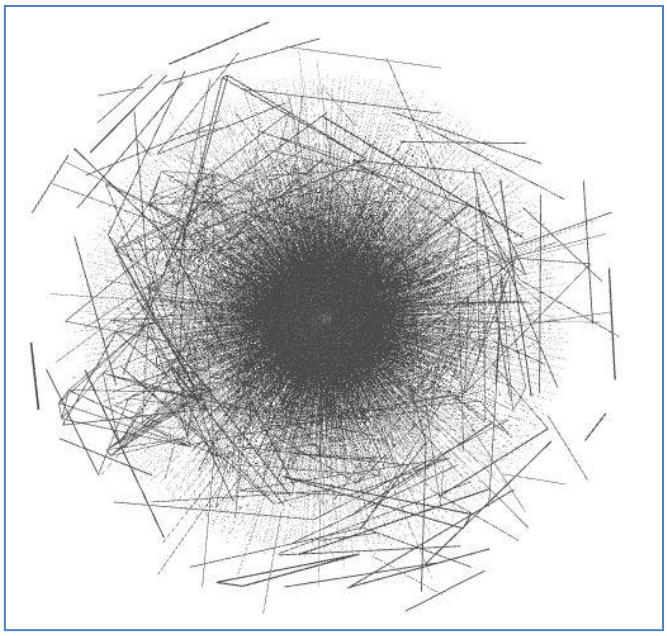

Ilustração 7 - Rede de artigos cluster MEDHIGH no bimestre 1

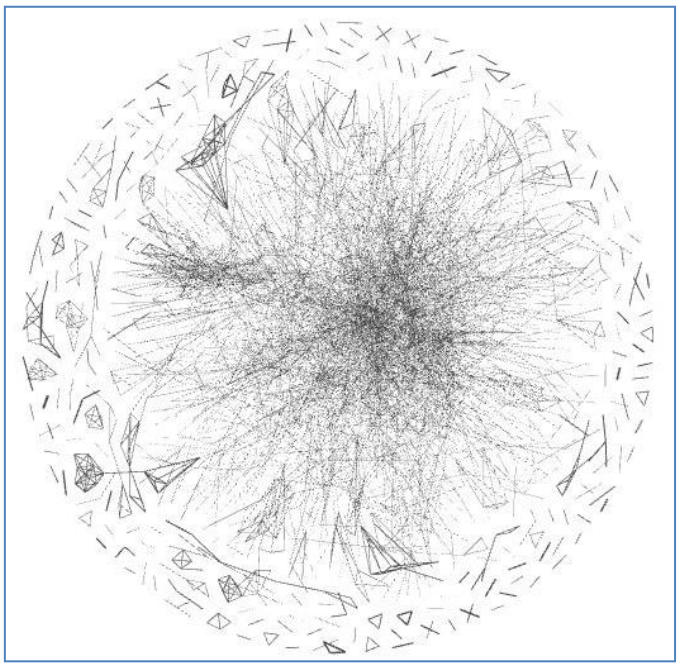

Ilustração 8 - Rede de artigos do cluster MEDHIGH no bimestre 1 após retirada top 200 vendas

A estrutura de rede do cluster MEDHIGH sem retirada dos *top* 200, ao contrário do que vimos no cluster BIG, revela já a existência de alguns componentes sem ligações ao núcleo constituído pelos artigos de elevada frequência. No entanto apresenta um grau médio semelhante (41) e uma menor modularidade (0,479).

Após a retirada dos *top* 200, obtivemos uma rede com menos cerca de 60% dos nós (3896 vs 1613: menos 59%) e 5% das ligações (80027 vs 4003).

Graficamente (ver ilustração 8), obtivemos uma segunda rede praticamente sem núcleo, e com uma estrutura em comunidades patente na modularidade (0,876).

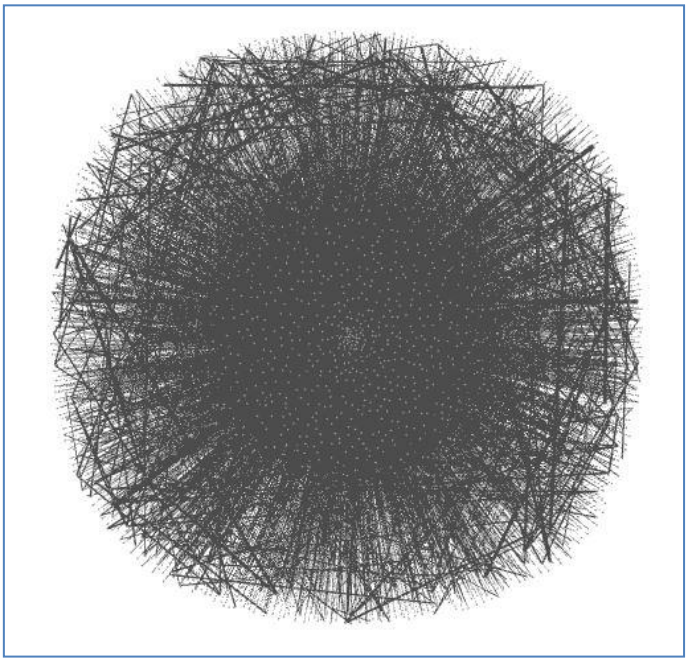

Ilustração 9 - Rede de artigos cluster MEDLOW no bimestre 1

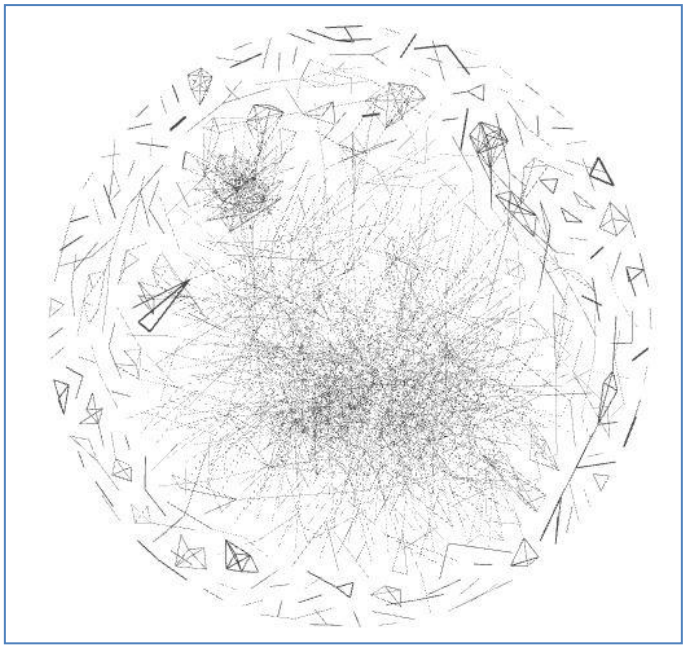

Ilustração 10 - Rede de artigos do cluster MEDLOW no bimestre 1 após retirada top 200 vendas

À semelhança dos *cluster* anteriores, as redes obtidas no cluster MEDLOW caracterizam-se por uma descida acentuada no número de nós (2832 vs 1065: menos 62%) e de ligações (55051 vs 2028: menos 96%), e um aumento acentuado na modularidade (0,456 vs 0,901).

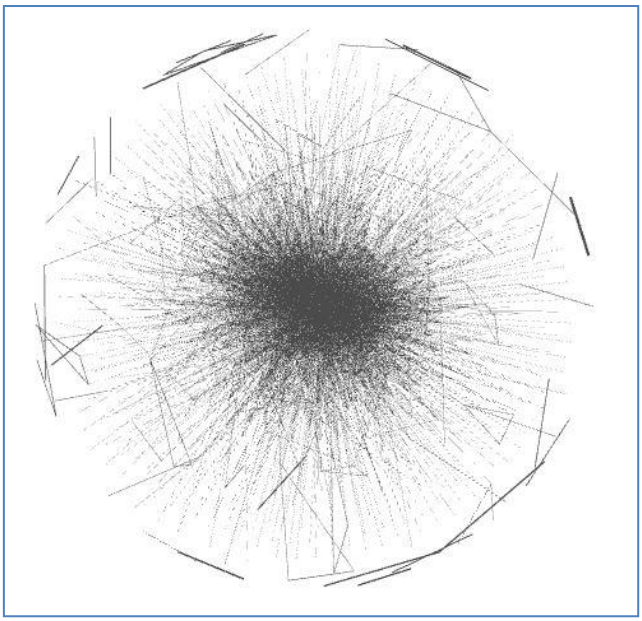

Ilustração 11 - Rede de artigos cluster SMALL no bimestre 1

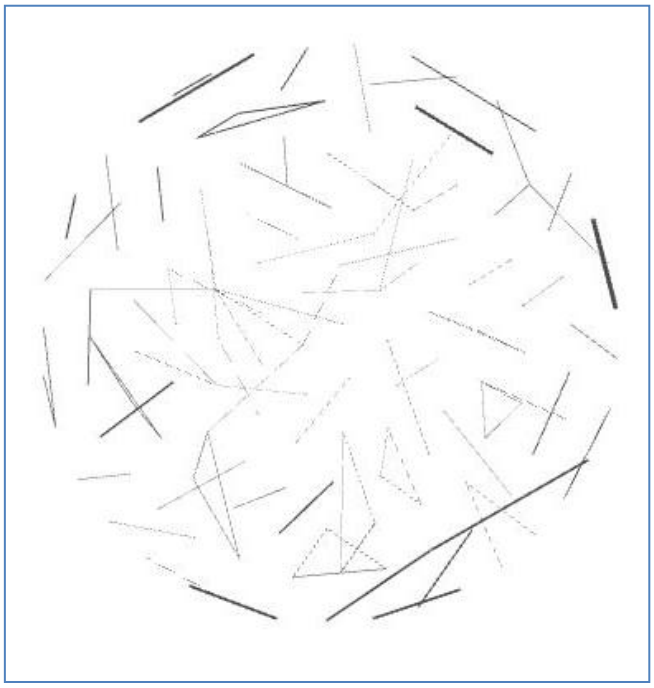

Ilustração 12 - Rede de artigos do cluster SMALL no bimestre 1 após retirada top 200 vendas

Nas redes obtidas no cluster SMALL, é evidente a sua diferença face aos clusters anteriores, logo na rede original, patente num grau médio muito inferior (apenas 13,8) e que graficamente se reflete numa rede menos densa e menos concentrada.

Após a retirada dos *top* 200 ficamos com uma rede em componentes pequenos e isolados.

Concluímos assim que ao retirar artigos top vendas, reduzimos consideravelmente a complexidade das redes a analisar, pelo que bastará apenas retirar os 100 *top* vendas para obter redes mais fáceis de construir e analisar.

Por outro lado, este efeito permitir-nos-á ampliar a análise ao ano completo, e, desta forma, incluirmos artigos com frequência muito reduzida (que de outra forma, seriam filtrados pelo suporte mínimo absoluto).

4.6 Redes de Artigos com base em Transações Comuns

4.6.1 Construção de Redes para Deteção de Comunidades (Cluster SMALL)

Tomando de base as redes construídas para cada *cluster*, após retirada dos top 100 vendas e resultantes de um ano completo de transações, seguiremos a metodologia indicada nos fundamentos teóricos, ligeiramente adaptada para incluir uma segunda variante que será a retirada de mais artigos top vendas, utilizando o cluster menor (SMALL).

Desta forma, recorrendo a script no R (ver anexo 2), são testados vários cenários, seguindo várias alternativas para cada uma das variantes:

- suporte mínimo relativo : - de 0,01% a 0,05%, em passos de 0,01%

- top máximo de vendas : - de 100 a 300, em passos de 10

Obtivemos desta forma vários cenários, nos quais pretende-se maximizar a modularidade das redes resultantes, sem esquecer que pretendemos redes que permitam obter comunidades de artigos com reduzida frequência.

Podemos ver nas ilustrações 13 e 14, que o suporte mínimo relativo de 0,01% domina todos os outros níveis de suporte em termos de tamanho de rede. Por outro lado, mantendo fixo este nível de suporte, vemos que a modularidade é crescente, mas atinge um *plateau* para o valor de top máximo de 230.

Assim, utilizaremos estes dois valores para o cálculo da rede que servirá de base para a deteção de comunidades.

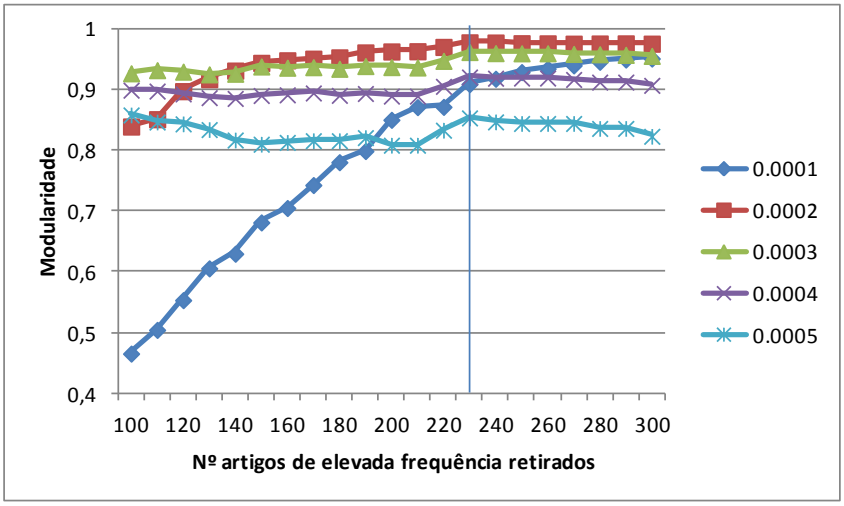

Ilustração 13 - Variação da modularidade para cenários de suporte mínimo e top máximo de vendas - Cluster SMALL

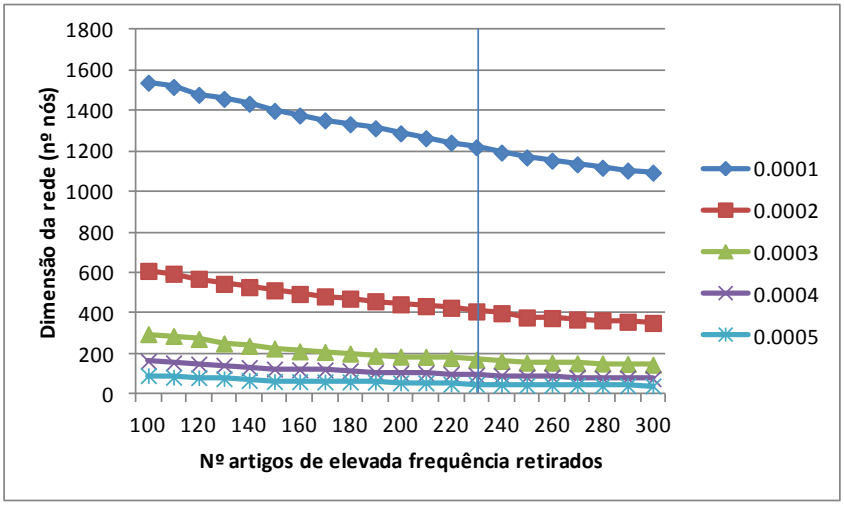

Ilustração 14 - Variação do tamanho de rede para cenários de suporte mínimo e top máximo de vendas - Cluster SMALL

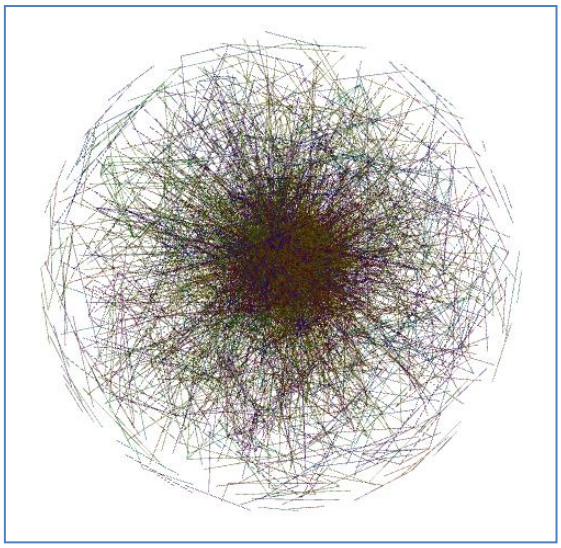

Ilustração 15 - Rede obtida com top máximo 100 e suporte relativo superior a 0,0001 - Cluster SMALL

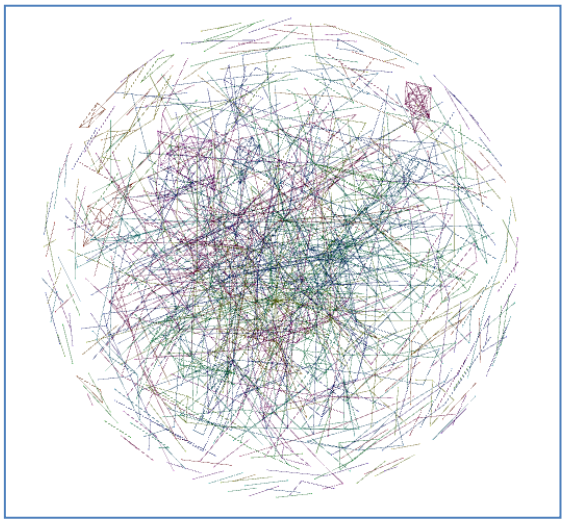

Ilustração 16 - Rede obtida com top máximo 230 e suporte relativo superior a 0,0001 - Cluster SMALL

Como podemos ver nas Ilustrações 15 e 16 (ver acima), a exclusão dos 230 artigos com maior frequência alterou significativamente a estrutura e a dimensão da rede de artigos (de 5004 ligações entre 1538 nós para 1482 ligações entre 1202 nós), eliminando-se a área mais densa que era polarizada pelos artigos top vendas retirados (densidade da rede diminuiu de 0,004 para 0,002).

Nº de nós	top 100	top 230
Mais de 100	2	1
Entre 50 e 100	2	2
Entre 10 e 50	7	18
Entre 5 e 10	25	22
Menos de 5	191	190
Total	**227**	**233**

Tabela 9 - Comparativo entre Top100 e Top230 em termos de comunidades - cluster SMALL

Podemos ver pela tabela 9 que apesar de se ter perdido uma comunidade de dimensão superior a 100, o número de comunidades aumentou, devido sobretudo ao aparecimento de um número superior de comunidades entre 10 e 50 artigos.

Através da tabela incluída no anexo 3, que representa as transferências havidas a partir das 6 maiores comunidades obtidas na rede com o top100, constatámos que as maiores comunidades são repartidas em comunidades mais pequenas. Por outro lado, dos 852 artigos representados nestas 6 comunidades, 319 (37,4%) não se encontram representados na rede top230.

4.6.2 Construção de Redes para Deteção de Comunidades (Cluster BIG)

No caso do cluster BIG, foi necessário espaçar os pontos de corte de top máximo de vendas para encontrar o plateau de modularidade que permitisse conciliar com redes de maior dimensão (ver Ilustrações 17 e 18).

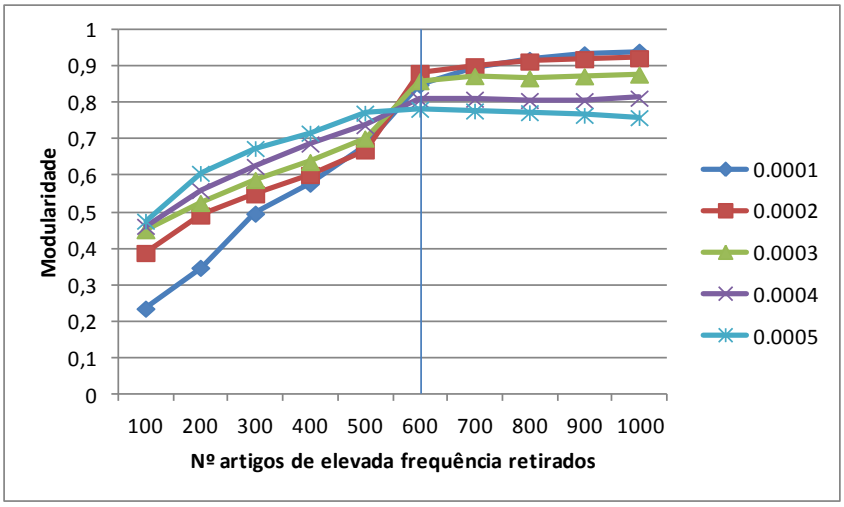

Ilustração 17 - Variação da modularidade para cenários de suporte mínimo e top máximo de vendas - Cluster BIG

Escolhemos para pontos de corte um suporte mínimo relativo de 0.01% e um top máximo de vendas de 600.

Embora um suporte de 0,02% permitisse um valor mais elevado de modularidade, obteríamos uma rede de menor dimensão, pelo que se optou por 0,01%.

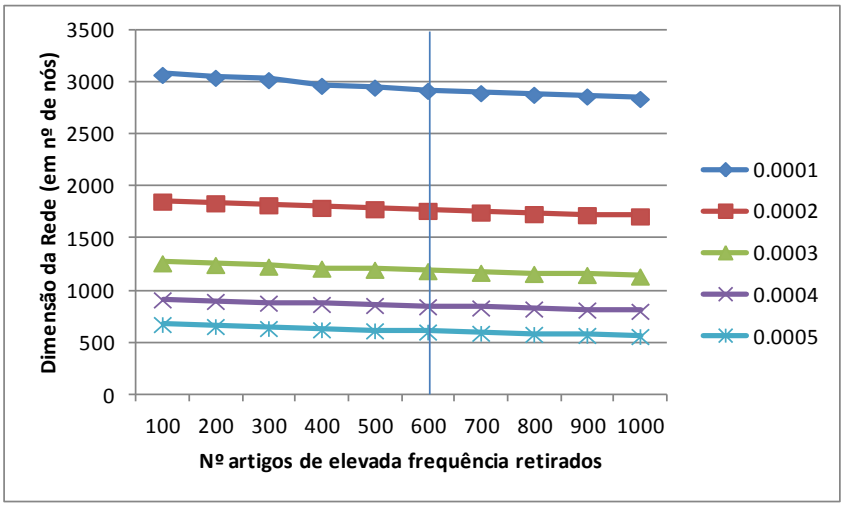

Ilustração 18 - Variação do tamanho de rede para cenários de suporte mínimo e top máximo de vendas - Cluster BIG

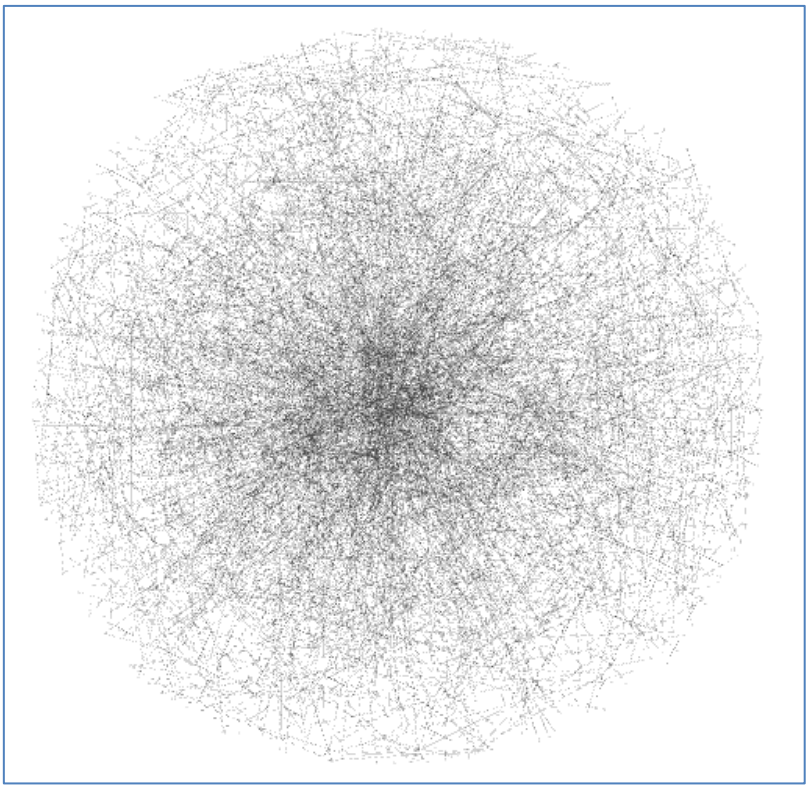

Ilustração 19 - Rede obtida com top máximo 600 e sup min 0,0001 - Cluster BIG

4.6.3 Cálculo do Valor das Comunidades (Cluster SMALL)

Após aplicar o algoritmo de deteção de comunidades no R (ver anexo 2), encontrámos 233 comunidades, distribuídas pelo valor de utilidade calculado conforme ilustração 8.

Podemos ver que a grande maioria das comunidades (84,5%) tem utilidade abaixo de 0,6. Visto que níveis superiores de utilidade indicam uma relação maior entre os artigos, vamos caracterizar duas comunidades de utilidade superior a 1.

Para além da utilidade, e dado termos como objetivo a obtenção de comunidade de artigos de baixa frequência, optámos por construir uma nova métrica para a avaliação de comunidades. Esta métrica relaciona-se com os valores de *lift* presentes nas ligações entre membros da comunidade e expressa-se por:

$$LIFT\ Médio\ (corrigido) = \sum_{(p1,p2)\in E_i} \max\left(300, \frac{P(p1|p2)}{P(p1)}\right) / \sum_{(p1,p2)\in E_i} 1$$

Equação 6 - Cálculo do lift médio corrigido

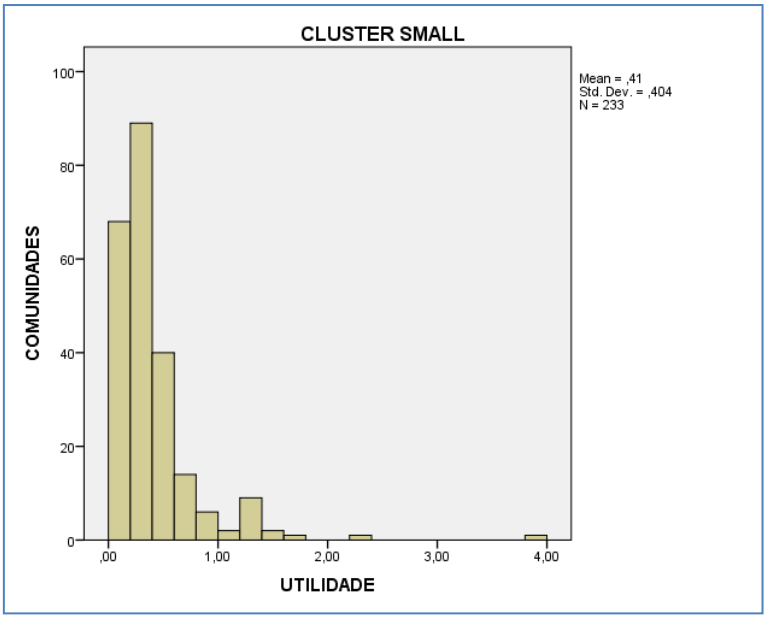

Ilustração 20 - Distribuição das comunidades por nível de utilidade - Cluster SMALL

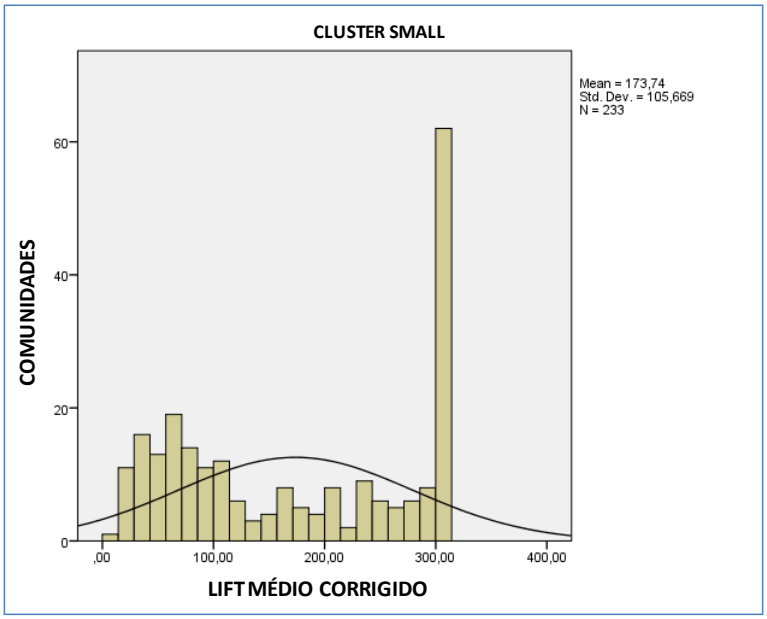

Ilustração 21 - Distribuição das comunidades por *lift* corrigido - Cluster SMALL

Ao contrário da confiança, o valor de *lift* não têm limite superior, pelo que valores muito altos poderiam tornar a comparação entre as comunidades mais difícil. Estabeleceu-se então um limite de *lift* por ligação de 300. Pela ilustração 21 podemos ver que um número significativo de comunidades atinge o máximo, o que é demonstrativo da força da ligação entre os seus elementos.

4.6.4 Detalhe de algumas comunidades - cluster SMALL

4.6.5 COM 31 - 7 ARTS - UTIL 3,92 - LIFT CORR 173

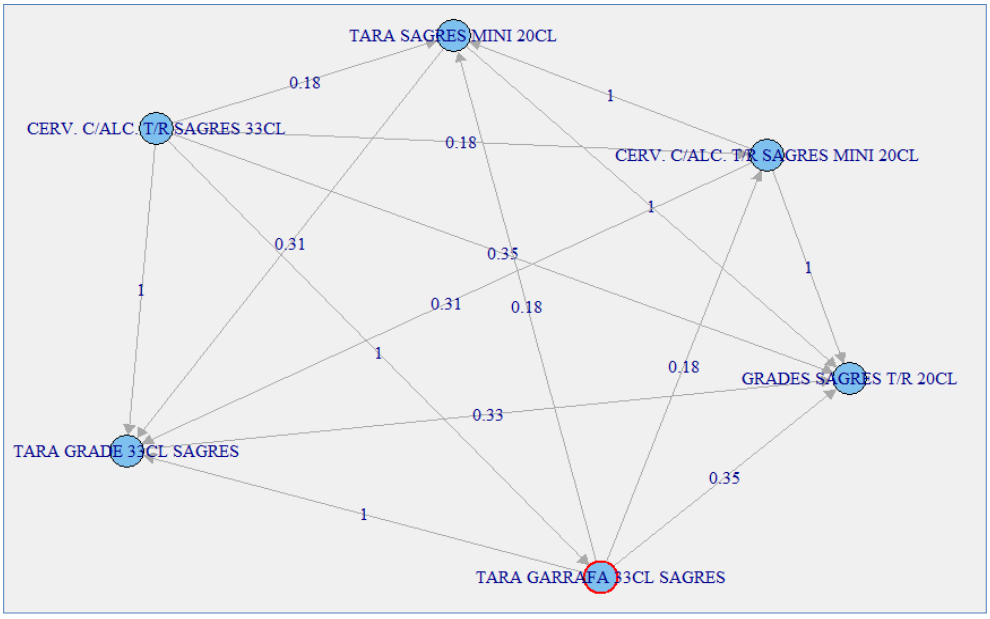

Ilustração 22 - Detalhe da Comunidade 31 (Confiança em label) - Cluster SMALL

SKU A	SKU B	P(A/B)	P(B/A)	LIFT_A	LIFT_B
CERV. C/ALC. T/R SAGRES 33CL	TARA GARRAFA 33CL SAGRES	1	1	804	804
CERV. C/ALC. T/R SAGRES 33CL	TARA SAGRES MINI 20CL	0,18	0,17	142	121
CERV. C/ALC. T/R SAGRES 33CL	CERV. C/ALC. T/R SAGRES MINI 20CL	0,18	0,17	142	121
CERV. C/ALC. T/R SAGRES 33CL	TARA GRADE 33CL SAGRES	1	0,46	804	118
CERV. C/ALC. T/R SAGRES 33CL	GRADES SAGRES T/R 20CL	0,35	0,15	278	39
TARA GARRAFA 33CL SAGRES	TARA SAGRES MINI 20CL	0,18	0,17	142	121
TARA GARRAFA 33CL SAGRES	CERV. C/ALC. T/R SAGRES MINI 20CL	0,18	0,17	142	121
TARA GARRAFA 33CL SAGRES	TARA GRADE 33CL SAGRES	1	0,46	804	118
TARA GARRAFA 33CL SAGRES	GRADES SAGRES T/R 20CL	0,35	0,15	278	39
TARA SAGRES MINI 20CL	GRADES SAGRES T/R 20CL	1	0,43	692	115
CERV. C/ALC. T/R SAGRES MINI 20CL	TARA SAGRES MINI 20CL	1	1	692	692
CERV. C/ALC. T/R SAGRES MINI 20CL	GRADES SAGRES T/R 20CL	1	0,43	692	115
TARA SAGRES MINI 20CL	TARA GRADE 33CL SAGRES	0,31	0,14	212	37
CERV. C/ALC. T/R SAGRES MINI 20CL	TARA GRADE 33CL SAGRES	0,31	0,14	212	37
TARA GRADE 33CL SAGRES	GRADES SAGRES T/R 20CL	0,33	0,31	85	82

Tabela 10 - Níveis de confiança das regras encontradas na comunidade 31 - Cluster SMALL

Nesta comunidade (ver ilustração 22) aparecem apenas artigos relacionados com uma marca de cervejas, mas podemos ver que existem relações com níveis de confiança muito elevados. Podemos dar como exemplo que em todas as transações em que apareceu o item TARA GARRAFA 33CL SAGRES, o item CERV. C/ALC. T/R SAGRES 33CL foi também registado. Trata-se de dois artigos expetavelmente relacionados, visto serem garrafas de cerveja de tara retornável e a devolução da respetiva tara.

Apesar de existirem níveis de confiança abaixo de 20%, como se tratam de artigos de muito reduzida frequência, equivalem a níveis de lift muito elevados, como podemos ver na tabela 10.

4.6.6 COM 39 - 4 ARTS - UTIL 1,6 - LIFT CORR 300

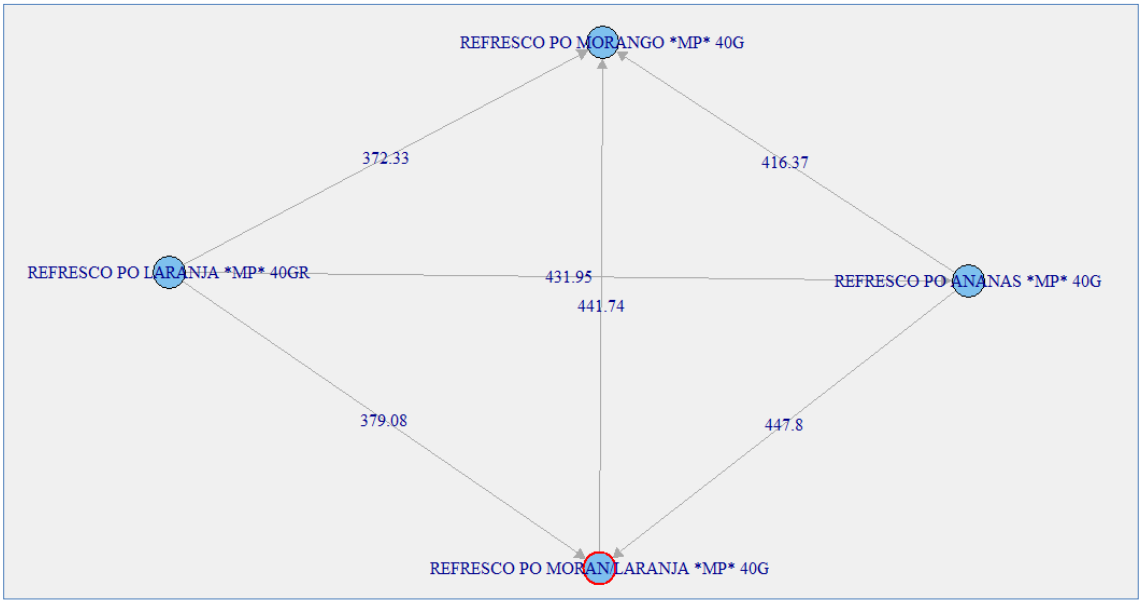

Ilustração 23 - Detalhe da Comunidade 39 (Lift em Label)- Cluster SMALL

SKU A	SKU B	P(A/B)	P(B/A)	LIFT_A	LIFT_B
REFRESCO PO LARANJA *MP* 40GR	REFRESCO PO MORAN/LARANJA *MP* 4	0,4	0,34	379	363
REFRESCO PO LARANJA *MP* 40GR	REFRESCO PO MORANGO *MP* 40G	0,4	0,2	372	352
REFRESCO PO LARANJA *MP* 40GR	REFRESCO PO ANANAS *MP* 40G	0,46	0,27	432	432
REFRESCO PO MORAN/LARANJA *MP* 4	REFRESCO PO MORANGO *MP* 40G	0,42	0,25	442	436
REFRESCO PO ANANAS *MP* 40G	REFRESCO PO MORAN/LARANJA *MP* 4	0,28	0,4	448	429
REFRESCO PO ANANAS *MP* 40G	REFRESCO PO MORANGO *MP* 40G	0,26	0,23	416	394

Tabela 11 - Níveis de confiança das regras encontradas na comunidade 39 - Cluster SMALL

Através da ilustração 23, podemos ver as relações existentes entre os vários sabores de refresco em pó. Embora os níveis de confiança entre os vários artigos se situem entre

0,2 e 0,46, quando analisamos os valores de *lift* (ver tabela 11) verificamos que os níveis de associação entre este conjunto de artigos são muito elevados. A baixa frequência esperada de qualquer um destes produtos implica que o nível de confiança não é uma boa medida dos níveis de associação.

4.7 Redes de Artigos com base em Clientes Comuns

As redes de artigos com base em clientes comuns permitem obter relações de compra fora da restrição da transação, mas apenas são possíveis de construir caso exista registo do cliente em número significativo de transações, como se trata deste caso.

Para eliminar ligações pouco representativas, estabelecemos como limite mínimo absoluto a existência de pelo menos 10 clientes comuns a dois artigos durante um ano. Este limite traduz-se num suporte mínimo relativo superior, pelo que iremos utilizar como patamar mínimo, para a análise de sensibilidade, um suporte relativo de 0,1%.

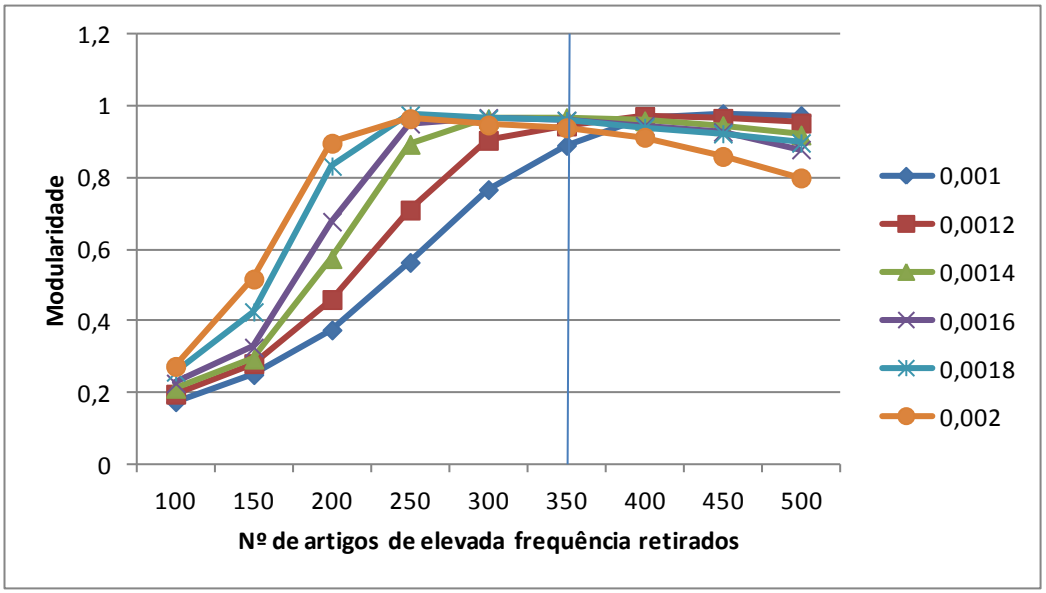

Ilustração 24 - Variação da modularidade para cenários de suporte mínimo e top máximo de vendas - Cluster SMALL

4.7.1 Construção de Redes para Deteção de Comunidades (Cluster SMALL)

De forma análoga ao ponto anterior, a escolha de um cenário de suporte mínimo relativo e top máximo de vendas, depende da modularidade, que se pretende maximizar, mas também de garantir que o tamanho de rede seja suficiente para o aparecimento de regras que contemplem artigos de frequência mais baixa.

Através das ilustrações 24 e 25 vemos que a escolha deverá recair num suporte mínimo relativo de 0,1%, com um top máximo de 350. Esta escolha garante-nos uma modularidade de cerca 0,9, ao ajustarmos a rede através dos artigos de elevada frequência e não através da base inferior de frequências. Outra alternativa, com um

valor aproximado de modularidade, seria um suporte mínimo de 0,12% e um top máximo de 300, mas resultaria um rede mais pequena (386 vs 481) e o ajustamento seria efetuado sobretudo através do suporte mínimo, o que prejudicaria o nosso objetivo de nos centrar na cauda longa.

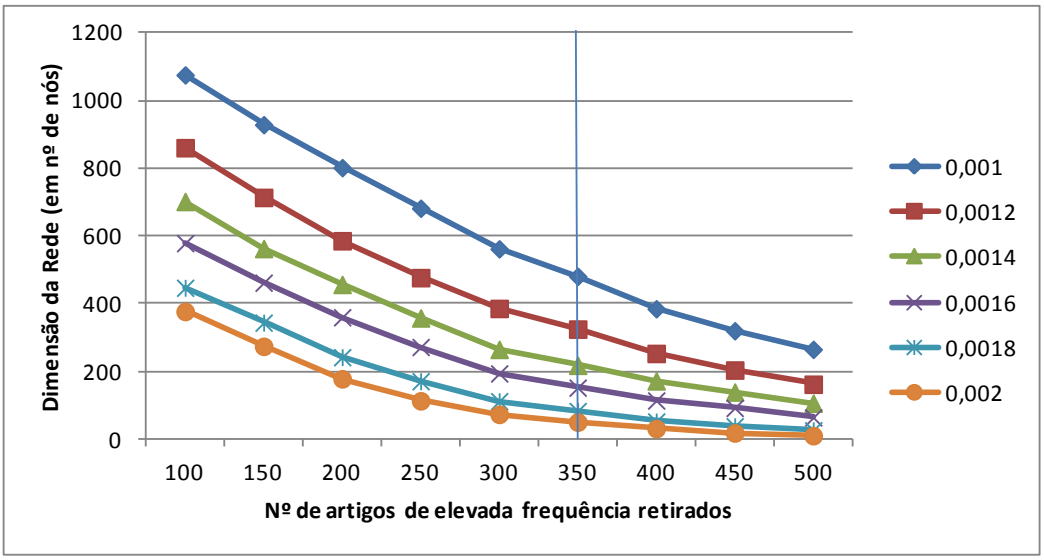

Ilustração 25 - Variação do tamanho de rede para cenários de suporte mínimo e top máximo de vendas - Cluster SMALL

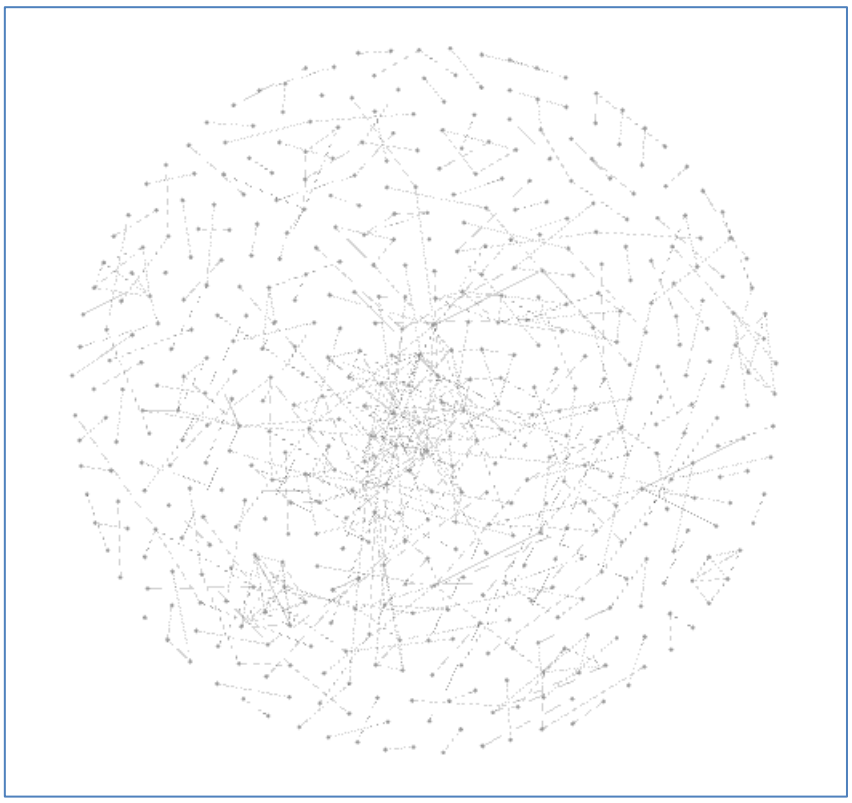

Ilustração 26 - Rede obtida com top máximo 350 e sup min 0,001 - Cluster SMALL

4.8 Redes de Clientes

Para a representação de clientes numa rede social, escolhemos como métrica de ligação a semelhança das categorias/tipo de marca compradas entre cada par de clientes. A construção desta métrica levou à necessidade de garantir informação suficiente de compras dos clientes para proceder a uma correta caracterização, o que obrigou aos seguintes passos de pré-processamento:

Numa primeira fase, eliminámos da análise todos os clientes que não tiveram pelo menos uma transação em cada um dos bimestres, ou que não tivessem comprado pelo menos em 100 unidades base/tipo de marca diferentes (para garantir que exista uma boa qualidade em termos de diversidade de compra), ou que não tivessem efetuado pelo menos 50 transações. Após esta fase, obtivemos uma amostra de cerca de 13000 clientes. Este filtro é necessário, pois se não existir informação suficiente ao nível do cliente, podemos incorrer em classificações erradas. Por outro lado, terá sempre de ser equacionado um equilíbrio entre os clientes que são retirados através deste filtro e a qualidade da informação suficiente para a análise, sob pena de reduzirmos exageradamente o número de clientes a analisar, ou, no outro extremo, obtermos resultados enviesados.

Numa segunda fase, coloca-se a questão de qual o nível de detalhe de produto que devemos considerar para este tipo de análise. Um nível muito elevado de detalhe poderá resultar em muitas repartições de produto com poucas compras de clientes, e por outro lado, perdermos alguma visão de conjunto. Outra questão relevante é a de colocarmos em perspetiva que tipo de marca estamos a considerar: se marca própria ou de fabricante, pois tem impacto nas escolhas de compra dos clientes. Após alguns testes optámos por agrupar os produtos ao nível da categoria, cruzado com tipo de marca.

Para garantir que não incluíamos agrupamentos de produtos com reduzida base de compra em termos de clientes, eliminámos todas as categorias/tipo de marca que não tivessem compras de pelo menos 1% dos clientes. De seguida calculámos a penetração de cada categoria/tipo de marca nas compras de cada cliente e selecionámos apenas as 150 com maior dispersão (medida pelo desvio-padrão). Esta redução foi essencial no nosso estudo, devido sobretudo a restrições de processamento. Este passo poderia ser evitado caso os recursos de processamento fossem mais substanciais.

A terceira fase tratou-se de elaborar scores para cada cliente com base em 10 fatores mais significativos encontrados através de análise componentes principais aplicada aos pesos que cada categoria/tp_marca têm nas compras de cada cliente (ver ilustração 27). Através da aplicação da análise em componentes principais obtivemos novas variáveis que, por não estarem correlacionadas entre si, acrescentam informação adicional, e explicam pelo menos metade da variância.

A quarta fase residiu na construção da matriz de dissemelhanças, através da distância euclidiana entre clientes com base nos scores obtidos na fase anterior.

Como resultado da quarta fase, obtivemos uma matriz de dissemelhanças com cerca de 85 milhões de pares. Como forma de restringir apenas a pares com um elevado nível de semelhança optámos por estabelecer uma distância euclidiana máxima de 3, o que reduziu a matriz a apenas 500 mil pares, correspondendo a 11842 clientes distintos.

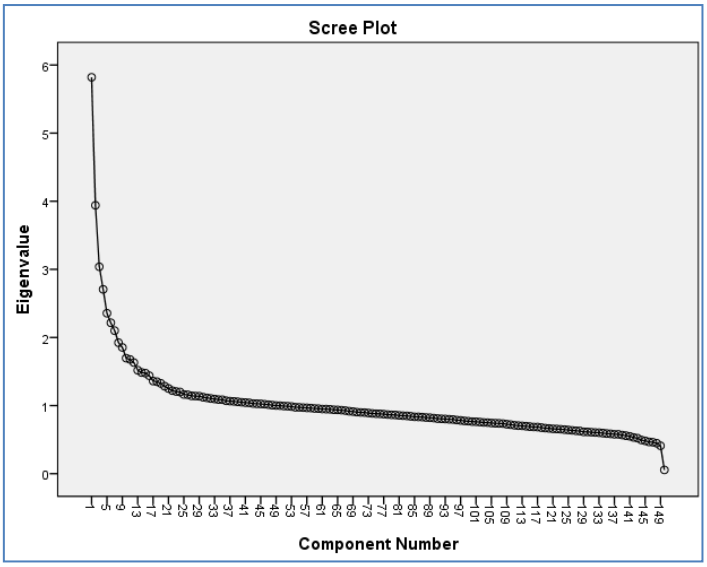

Ilustração 27 - Screeplot das componentes principais encontradas

Após importação das ligações para o Gephi, e alguns testes com a distância euclidiana máxima entre clientes, obtivemos uma rede muito esparsa (densidade de 0,007), com modularidade 0,404, que resultou em 58 comunidades, das quais apenas 10 de dimensão relevante (ver ilustração 28).

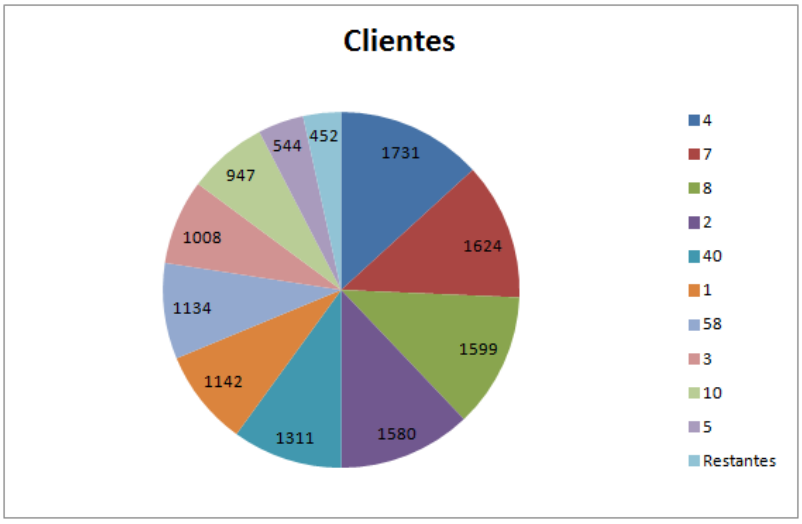

Ilustração 28 - Dimensão das comunidades encontradas

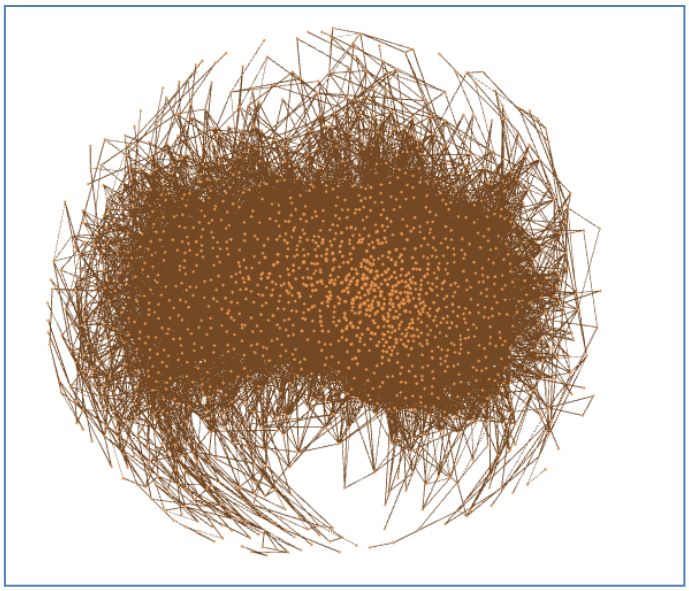

Ilustração 29 - Detalhe da comunidade 4

Na ilustração 29 podemos ver uma representação da maior comunidade encontrada.

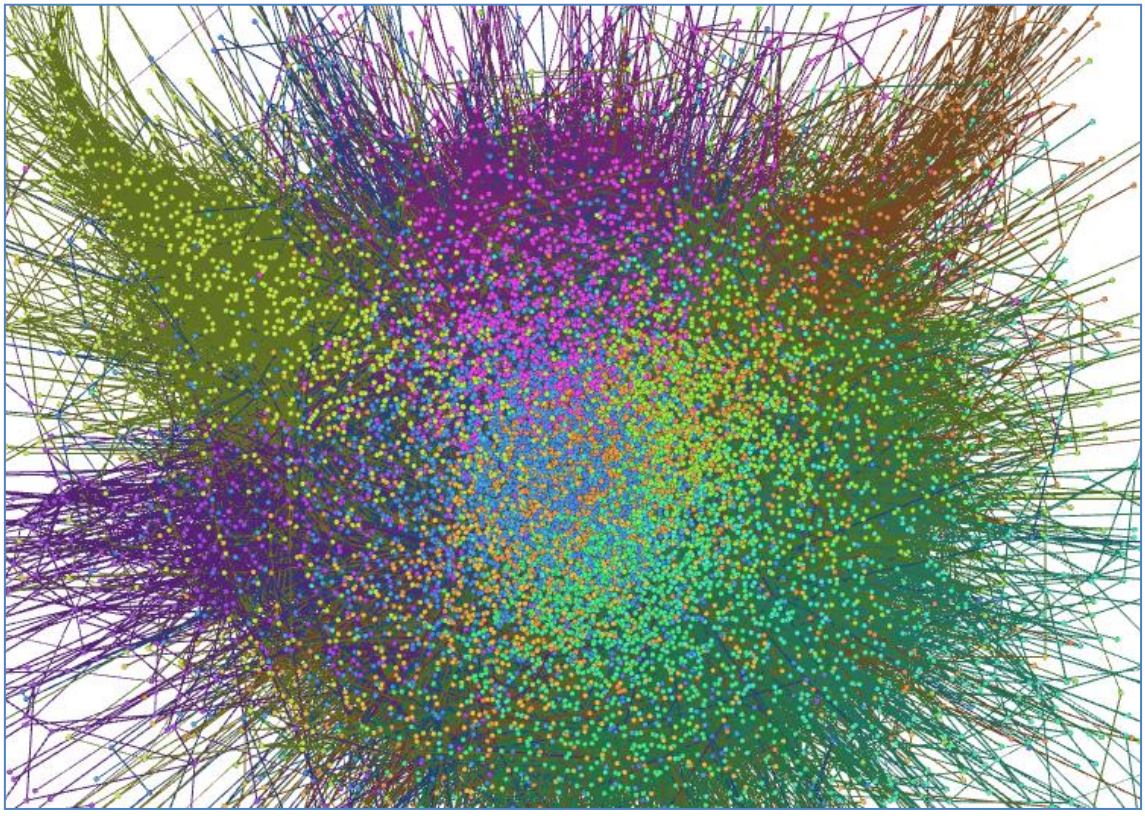

Ilustração 30 - Detalhe da rede de clientes obtida (com mapeamento de comunidades por cor)

Como podemos ver na ilustração 30, as comunidades encontradas na rede social, e principalmente no núcleo da rede, estão muito interligadas entre si.

As 10 comunidades foram caracterizadas através de quatro vertentes:

- quais os tipos de artigos que mais/menos compram em relação à média dos clientes

- qual o perfil de compra em termos de visita/valor de compra

- qual o seu posicionamento em relação às marcas

- qual o perfil sociodemográfico

		INDICES		
Comunidade	Nº Clientes	COMPRA POR CLIENTE	TRANSAÇÃO MÉDIA	Nº MÉDIO TRX
1	1142	105	114	92
2	1580	114	102	112
3	1008	78	71	110
4	1731	100	109	92
5	544	110	109	101
7	1624	87	91	96
8	1599	97	92	106
10	947	87	80	108
40	1311	117	133	88
58	1134	105	108	97
	12620			

Tabela 12 - Comportamento de Compra em cada comunidade de clientes

Podemos ver na tabela 12 que as diferentes comunidades têm comportamentos muito distintos em termos de compras. Destacamos a comunidade 3, que apresenta o valor de compra por cliente baixo, e uma transação média baixa. Por oposição, a comunidade 40 apresenta um valor elevado de compra por cliente, e também uma transação média elevada.

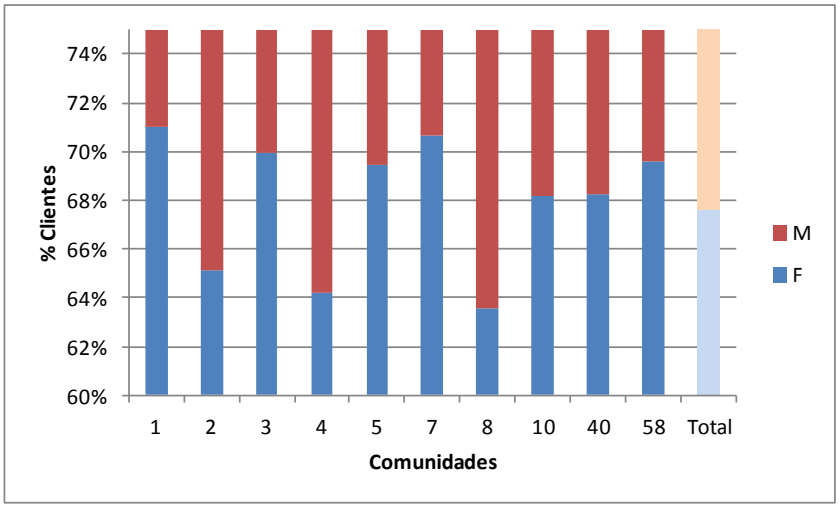

Ilustração 31 - Caracterização de cada comunidade em termos de género

Sabemos por conhecimento de negócio que o cartão de fidelização representa compras de uma família, pelo que o género indicado por referência a cada cartão oferece algumas

limitações em termos de análise. No entanto, reduzindo a informação obtida de cada comunidade a termos relativos, podemos ver na ilustração 31 que existem algumas diferenças a este nível. Destaca-se uma tendência para clientes do sexo masculino nas comunidades 2, 4 e 8.

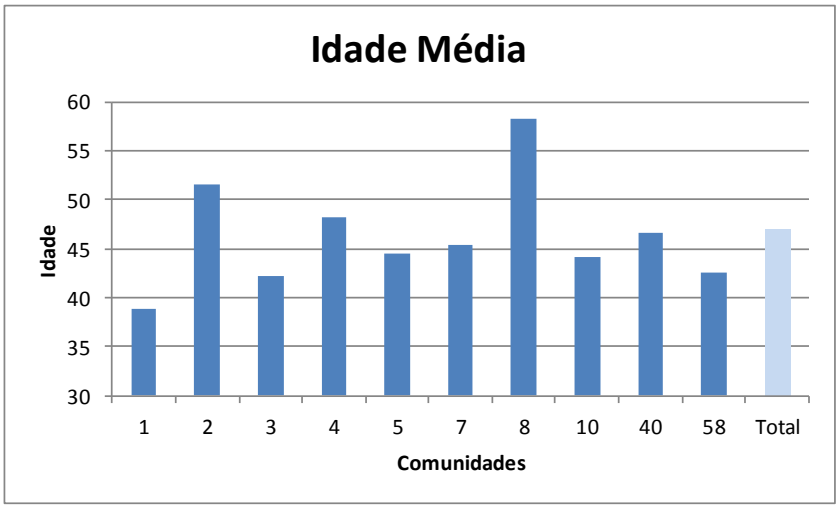

Ilustração 32 - Média etária de cada comunidade

Em termos de idade média de cada comunidade (ver ilustração 32), destacamos a elevada idade média da comunidade 8, por oposição, a uma comunidade mais jovem, a 1.

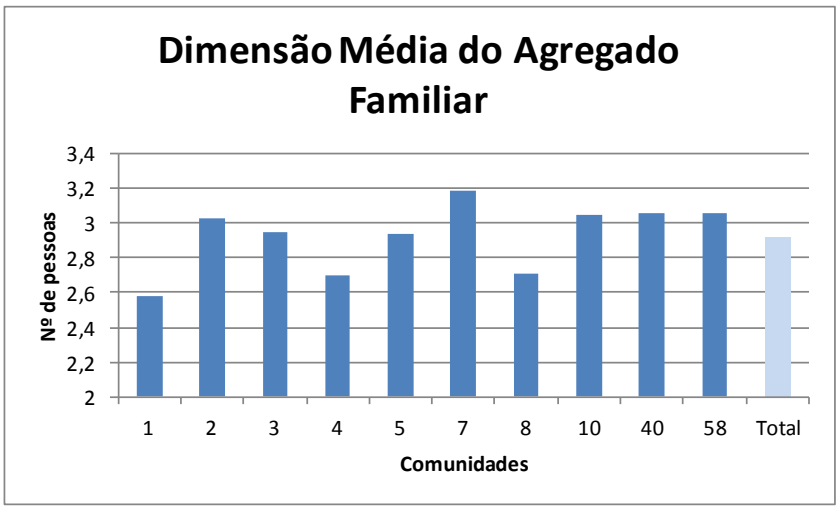

Ilustração 33 - Dimensão Média do Agregado Familiar por comunidade

Na Ilustração 33, constatamos que as comunidades 1, 4 e 8 se caracterizam por agregados familiares de menor dimensão. A comunidade 7 representa o índice mais elevado nesta variável.

Apresentamos em seguida o comportamento de compra de cada comunidade nas categorias principais, medido pela diferença relativamente à compra média, medida em índice. No eixo das coordenadas das ilustrações 34 a 43 está representada a diferença entre o índice de compra média na categoria da comunidade correspondente e o indíce de referência (100) relativo ao total dos clientes.

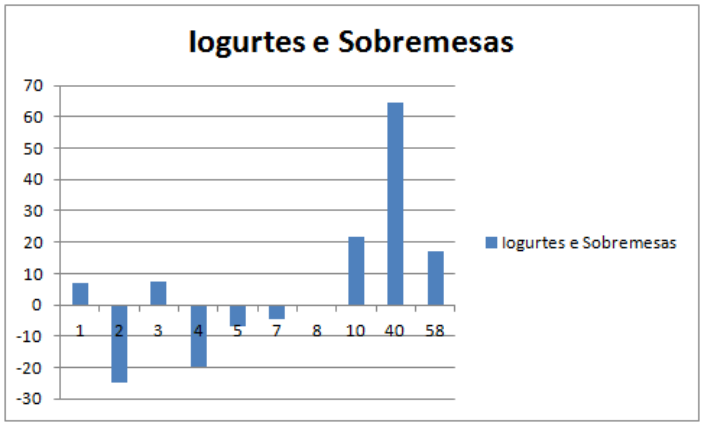

Ilustração 34 - Diferença face à compra média de Iogurtes e Sobremesas

Na ilustração 34 podemos ver que a comunidade 40 apresenta uma compra média em Iogurtes e Sobremesas cerca de 65% superior à média total. Por oposição, as comunidades 2 e 4 apresentam uma menor propensão á compra nesta categoria de artigos.

À semelhança da categoria de Iogurtes e Sobremesas, a comunidade 40 destaca-se pela elevada propensão (cerca de 63%) á compra na categoria de Leite e Bebidas de Soja.

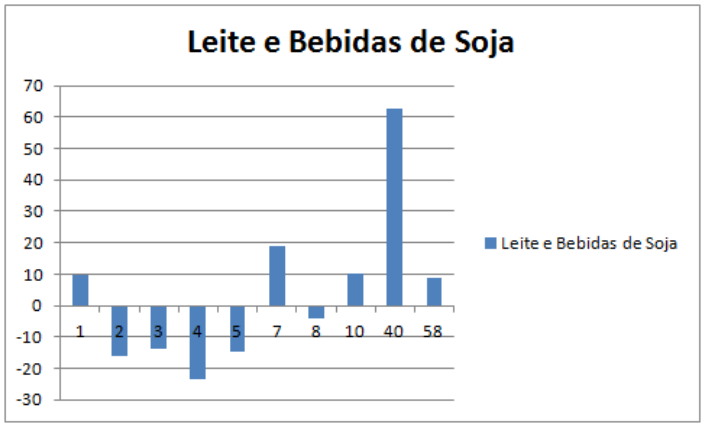

Ilustração 35 - Diferença face à compra média de Leite e Bebidas de Soja

Na categoria de Frutas, a comunidade 8 apresenta uma compra média 61% superior à média verificada no total de clientes (ver ilustração 36). De igual forma se destaca na compra de Peixe Fresco (ver ilustração 38).

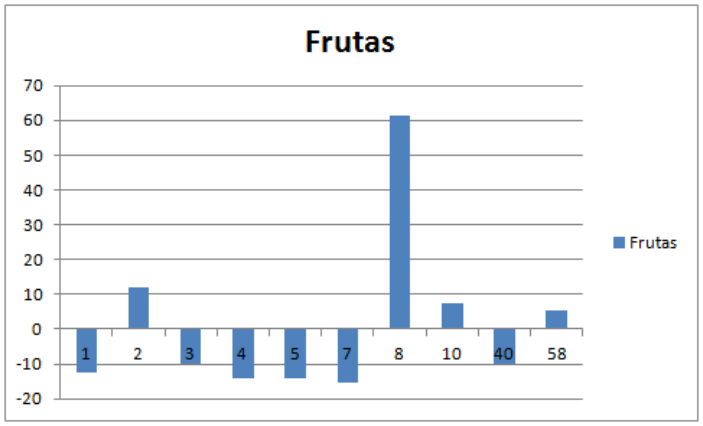

Ilustração 36 - Diferença face à compra média de Frutas

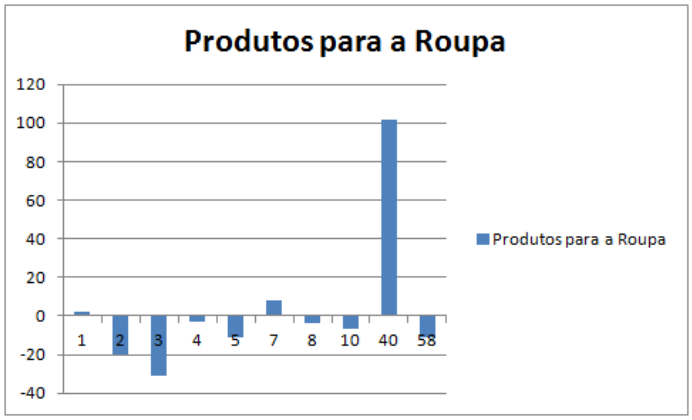

Ilustração 37 - Diferença face à compra média de Produtos para a Roupa

A comunidade 40 surge novamente em destaque na compra de Produtos para a Roupa, como podemos ver na ilustração 37.

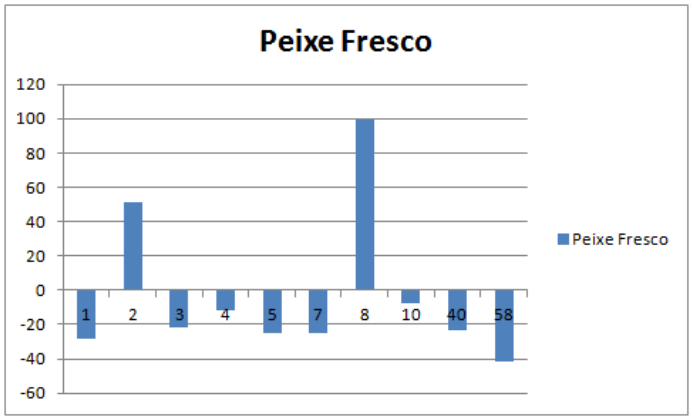

Ilustração 38 - Diferença face à compra média de Peixe Fresco

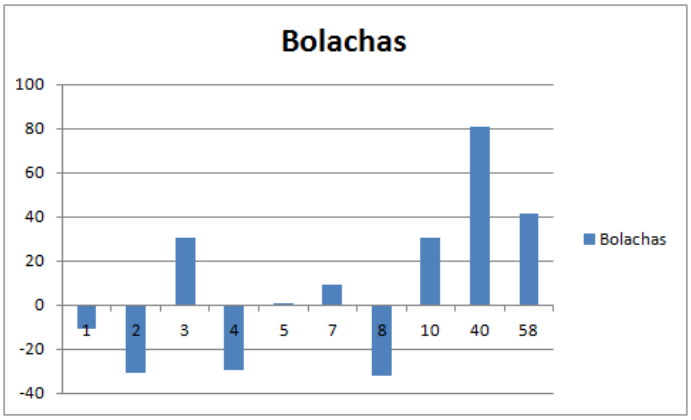

Ilustração 39 - Diferença face à compra média de Bolachas

Como podemos ver na ilustração 39, a compra de Bolachas é superior em cerca de 80% face à média na comunidade 40. As comunidades 10 e 58 também apresentam valores de compra superiores à média (30% e 41%, respetivamente).

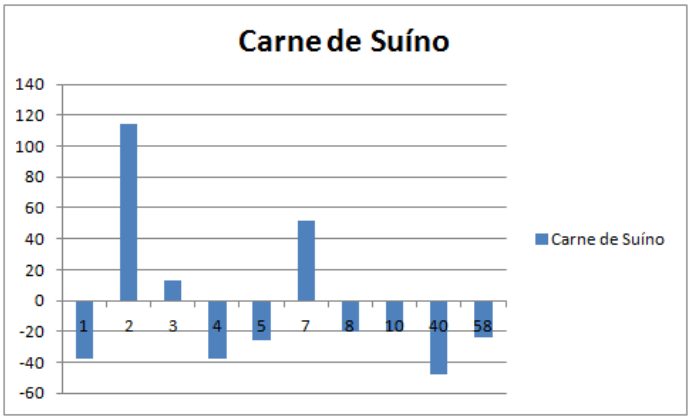

Ilustração 40 - Diferença face à compra média de Carne de Suíno

Na compra nas categorias de Aves e Suínos (ver ilustrações 40 e 41), destaca-se a comunidade 2, com valores de compra cerca do dobro da compra média nesta categoria.

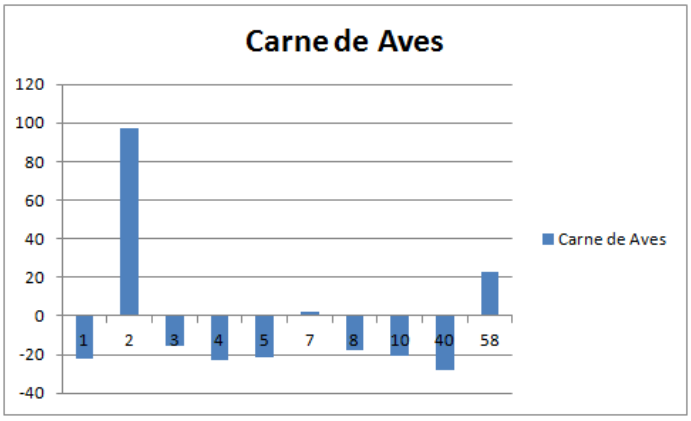

Ilustração 41 - Diferença face à compra média de Carne de Aves

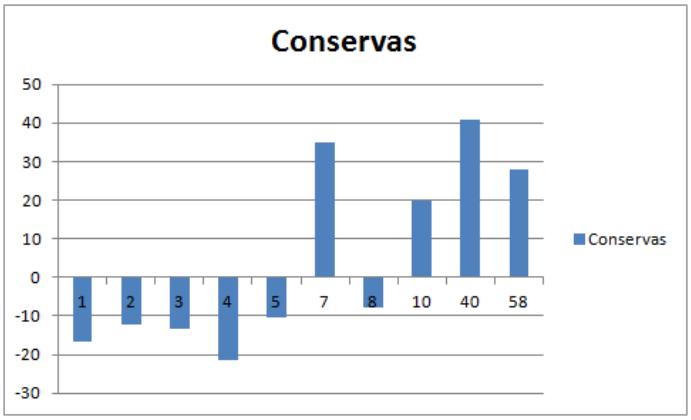

Ilustração 42 - Diferença face à compra média de Conservas

Na ilustração 42 destacam-se quatro comunidades pela compra superior à média na categoria de Conservas: 40, 7, 58 e 10.

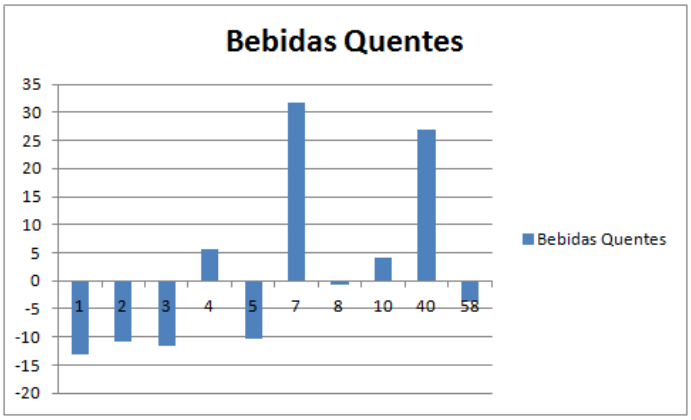

Ilustração 43 - Diferença face à compra média de Bebidas Quentes

As comunidades 7 e 40 apresentam um valor de compra média na categoria de Bebidas Quentes muito superior à média, como se pode ver na ilustração 43.

Nas tabelas 13 e 14 podemos ver um resumo das características mais marcantes de cada comunidade através das quatro vertentes atrás mencionadas. As tabelas refletem os aspetos mais relevantes obtidos através da análise dos resultados obtidos nas quatro vertentes: - sócio-demográfica, comportamento de compra, posicionamento face às marcas e compra nas categorias de artigos.

A título de exemplo, analisaremos a comunidade 40 (já evidenciada por diversas vezes nas análises anteriores. Podemos ver na tabela 13, que se trata de clientes mais novos, com um agregado familiar superior à média, que compra mais produtos da marcas de fornecedor, e com maior tendência (ver tabela 14) para comprar mais produtos de DPH (Detergentes e Produtos de Higiene), Lacticínios, Bolachas e Bebidas sem Alcool, em detrimento de produtos frescos (Talho e Legumes).

Num segundo exemplo, atente-se na comunidade 8. É o *cluster* que apresenta uma idade média superior, tendencialmente masculino, com um agregado familiar menor que a média, e que se destaca por um número médio de transações por cliente superior à média, refletindo um elevado número de visitas à loja (ver tabela 13). Procuram artigos de marca própria e categorias de produtos e tradicionais (ver tabela 14). Podemos considerar que se trata de um segmento sénior.

	SOCIO-DEMOGRAFICO	COMPORTAMENTO COMPRA	POSICIONAMENTO FACE ÁS MARCAS
1	O CLUSTER COM MENOS IDADE MÉDIA E AGREGADO FAMILIAR MÉDIO MAIS PEQUENO / MAIS FEMININO QUE A MÉDIA	VALOR MÉDIO DE TRANSAÇÃO SUPERIOR Á MÉDIA	COMPRA MENOS MARCA PRÓPRIA E PRIMEIRO PREÇO QUE A MÉDIA
2	MAIS IDADE, MAIS MASCULINO, MAIOR AGREGADO FAMILIAR QUE A MÉDIA	CLUSTER COM MAIOR VALOR DE COMPRA POR CLIENTE E MAIOR NÚMERO DE TRANSAÇÕES POR CLIENTE	COMPRA MAIS MARCA PRÓPRIA E PRIMEIRO PREÇO QUE A MÉDIA
3	MAIS NOVO, MAIS FEMININO QUE A MÉDIA	CLUSTER COM MENOR VALOR DE COMPRA POR CLIENTE E VALOR MÉDIO DE TRANSAÇÃO, MAS NÚMERO DE TRANSAÇÕES POR CLIENTE SUPERIOR À MÉDIA	COMPRA MAIS PRIMEIRO PREÇO QUE A MÉDIA
4	MAIS MASCULINO, MENOR AGREGADO FAMILIAR QUE A MÉDIA	VALOR MÉDIO DE TRANSAÇÃO SUPERIOR Á MÉDIA, MAS NÚMERO DE TRANSAÇÕES POR CLIENTE INFERIOR Á MÉDIA	COMPRA MENOS PRIMEIRO PREÇO E MARCA PRÓPRIA QUE A MÉDIA
5	-	COMPRA POR CLIENTE E VALOR MÉDIO DE TRANSAÇÃO SUPERIOR Á MÉDIA	CAMPEÃO DA MARCA EXCLUSIVA - COMPRA MENOS PP E MP QUE A MÉDIA
7	MAIS FEMININO QUE A MÉDIA - CLUSTER COM MAIOR AGREGADO FAMILIAR MÉDIO	CLUSTER COM MENOR VALOR MÉDIO DE COMPRA POR CLIENTE - TRANSAÇÃO MÉDIA E Nº DE TRANSAÇÕES POR CLIENTE INFERIOR Á MÉDIA	CAMPEÃO DO PRIMEIRO PREÇO
8	O CLUSTER COM IDADE MÉDIA MAIS ALTA - MAIS MASCULINO, MENOR AGREGADO FAMILIAR QUE A MÉDIA	Nº MÉDIO DE TRANSAÇÕES POR CLIENTE SUPERIOR Á MÉDIA	COMPRA MAIS MARCA PRÓPRIA QUE A MÉDIA
10	MAIS NOVO E MAIOR AGREGADO FAMILIAR QUE A MÉDIA	MENOR COMPRA POR CLIENTE E VALOR MÉDIO DE TRANSAÇÃO, MAS MAIOR Nº DE TRANSAÇÕES POR CLIENTE	COMPRA MAIS MARCA PRÓPRIA QUE A MÉDIA
40	MAIS NOVO E MAIOR AGREGADO FAMILIAR QUE A MÉDIA	CLUSTER COM MENOR VALOR DE COMPRA POR CLIENTE E VALOR MÉDIO DE TRANSAÇÃO, MAS NÚMERO DE TRANSAÇÕES POR CLIENTE SUPERIOR À MÉDIA	CAMPEÃO DAS MARCAS DE FORNECEDOR
58	MAIS NOVO QUE A MÉDIA	CLUSTER COM MENOR VALOR DE COMPRA POR CLIENTE E VALOR MÉDIO DE TRANSAÇÃO, MAS NÚMERO DE TRANSAÇÕES POR CLIENTE SUPERIOR À MÉDIA	COMPRA MAIS MARCA PRÓPRIA DO QUE A MÉDIA

Tabela 13 - Caracterização dos clusters através dos dados sócio-demográficos, comportamento de compra e posicionamento face á marca

	CATEGORIAS	
	+	-
1	ALIMENTAÇÃO INFANTIL - HIGIENE E PROTEÇÃO BEBÉ - BRINQUEDOS	BEBIDAS COM ALCOOL - PETCARE - PEIXE
2	TALHO - PEIXARIA	ALIMENTAÇÃO INFANTIL - HIGIENE E PROTEÇÃO BEBÉ - BRINQUEDOS -PETCARE - DIETÉTICOS - ARTIGOS CULTURA
3	TAKE-AWAY - PADARIA - COZINHA FÁCIL - DOÇARIA - BOLACHAS	BEBIDAS COM ALCOOL - PROD LOIÇA - INGREDIENTES BÁSICOS - PROD ROUPA
4	BEBIDAS COM ALCOOL - PETCARE - CERVEJA - PERFUMARIA E COSMÉTICA - DIETÉTICOS	ALIMENTAÇÃO INFANTIL - HIGIENE E PROTEÇÃO BEBÉ - CARNES ATENDIMENTO - COZINHA FACIL
5	TAKE-AWAY - BRINQUEDOS - DIETÉTICOS - ARTIGOS DE CULTURA	ALIMENTAÇÃO INFANTIL - HIGIENE E PROTEÇÃO BEBÉ - CARNES ATENDIMENTO
7	MERCEARIA - BACALHAU	ALIMENTAÇÃO INFANTIL - HIGIENE E PROTEÇÃO BEBÉ - BRINQUEDOS
8	BACALHAU - PEIXE FRESCO E CONGELADO - FRUTAS E LEGUMES - DIETÉTICOS	ALIMENTAÇÃO INFANTIL - HIGIENE E PROTEÇÃO BEBÉ - BRINQUEDOS - ARTIGOS DE CULTURA - BOLACHAS - APERITIVOS - CERVEJAS
10	SUMOS E REFRIGERANTES - BOLACHAS E PADARIA	ALIMENTAÇÃO INFANTIL - HIGIENE E PROTEÇÃO BEBÉ - BRINQUEDOS - PETCARE
40	DPH - BEBIDAS SEM ALCOOL - BOLACHAS - LACTICÍNIOS	TALHO - LEGUMES
58	LACTICINIOS E CONGELADOS - MERCEARIA	ALIMENTAÇÃO INFANTIL - HIGIENE E PROTEÇÃO BEBÉ - BEBIDAS COM ALCOOL

Tabela 14 - Caracterização dos clusters através do seu comportamento de compra nas categorias de produtos

5 Conclusões e Futuros Desenvolvimentos

O potencial da aplicação de análise de redes sociais à temática das relações entre artigos ou clientes ficou demonstrado ao longo do trabalho, quer na vertente analítica, como na de apresentação de resultados.

Partindo da aplicação da metodologia proposta por Chawla para construção de redes de artigos (Chawla et al, 2011), a base deste trabalho incidiu na sua adaptação para a exploração da "cauda longa" das vendas, através da retirada do impacto provocado pelos artigos de elevada rotação.

Ficou demonstrado ser possível através de análise de sensibilidade da modularidade e tamanho de rede aos parâmetros de retirada de artigos top vendas e suporte mínimo relativo, obter redes menos densas e mais focadas nos artigos de menor rotação. Por outro lado, esta estratégia reduz consideravelmente a exigência em termos de processamento, permitindo alargar a análise ao contexto multi-loja e, desta forma, obter uma maior profundidade de análise.

O agrupamento de lojas com base nas aproximações de gama de artigos utilizado permitiu que os contextos de loja fossem o mais aproximado possível dentro de cada *cluster* de lojas. Porém, existem especificidades regionais que poderiam ser incluídas ao agrupar lojas segundo clientes em comum.

A valoração da utilidade de comunidade através da métrica proposta - lift corrigido - revelou-se mais apropriada para comunidades de artigos de baixa rotação. Nestes casos, ainda que as relações existentes entre artigos da mesma comunidade apresentem, em alguns casos, níveis de confiança pouco elevados, a nova métrica permite contextualizá-las à luz da reduzida frequência destes artigos.

Por não dispormos de informação sobre as promoções realizadas nas lojas, ficamos limitados na informação obtida neste trabalho. Num mundo ideal, as vendas promocionais deveriam ser tratadas de forma especial, visto que uma relação entre artigos pode estar a ser construída com base numa promoção conjunta ou coincidente no tempo, e não refletir uma verdadeira e natural associação.

Na vertente de rede de clientes, apresentámos uma metodologia evoluída a partir da literatura já existente, que propõe um novo caminho a nível de segmentação baseada na aproximação entre perfis de compra.

A aplicação da deteção de comunidades à rede de clientes resultou em agrupamentos bem diferenciados, e por isso, fornece uma nova forma de segmentação de clientes por comportamento de compra.

5.1 Futuros Desenvolvimentos

As potencialidades da Análise de Redes Sociais podem ir além das aplicações focadas neste trabalho. Um dos possíveis desenvolvimentos seria a aplicação de redes bipartidas

(nós representando grupos de produtos e clientes) e a deteção de *overlaping communities*.

Existe a possibilidade de construir redes de lojas com base em clientes em comum, obtendo um mapeamento geográfico das lojas, mesmo que a sua localização não esteja disponível. Quando cruzado com o agrupamento de lojas por gama de artigos, podemos obter redes de artigos com contexto regional.

Outro caminho possível seria, a partir dos segmentos encontrados na rede de clientes, construir redes de artigos com base na afinidade, dentro de cada segmento. A lógica subjacente é a afinar as comunidades de artigos dentro de cada segmento de cliente, e, por comparação, obter novas informações que ajudem a caracterizar esses segmentos.

À semelhança do que hoje já se faz no comércio *online* (Oestreicher-Singer, Gal, 2013), poderá ser possível analisar o valor de um artigo com base no potencial de alavancar vendas de outros artigos. Um dos caminhos possíveis seria a aplicação do algoritmo *PageRank*.

Redes de clientes poderão servir também para a deteção de comportamentos fraudulentos (uso comercial do cartão, cartões duplicados), em que as relações entre os clientes poderão refletir o uso do mesmo NIF ou números de contacto idênticos.

Como se pode ver, o campo de aplicação de análise de redes sociais a uma empresa de retalho é extremamente vasto, pelo que será provavelmente uma ferramenta de uso crescente dentro deste domínio.

Por ser uma área que exige uma grande capacidade de processamento, será uma grande beneficiária dos avanços obtidos através das ferramentas de *big data*.

6 Referências Bibliográficas

Anderson, Chris (2006), *The Long Tail: Why the Future of Business is Selling Less of More*, Hyperion (2006), ISBN 978-1401302375

Newman, M. E. J. e Girvan , M. (2004), Finding and evaluating community structure in networks

Chawla, Nitesh V. e Raeder, Troy (2011), "Market basket analysis with networks" in *Social Network Analysis and Mining*, 2011, Vol 1, 2, pp 97-113

Chawla, Nitesh V. e Raeder, Troy (2009), "Modeling a Store's Product Space as a Social Network", 2009, pp 164-169

Chen, Yen-Liang et al (2005), "Market basket analysis in a multiple store environment" in *Decision Support Systems* 40 (2005) 339– 354, Elsevier

Fortunato, Santo (2010), "Community detection in graphs", Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino,Italy

Kim, Hyea Kyeong et al. (2012), "A product network analysis for extending the market basket analysis", in *Experts Systems with Applications*, Elsevier

Oliveira , Márcia e Gama , João (2012), "An overview of social network analysis" in *WIRE Data Mining Discov 2012*, 2: 99-115

Miguéis , V.L. (2012), "Customer data mining for lifestyle segmentation", in *Expert Systems with Applications*, 2012, Vol 39, 10, pp 9359-9366

Videla-Cavieres et al. (2013), "Extending market basket analysis with graph mining techniques: A real case", in *Expert Systems with Applications*, 2013, Vol 14, 4, pp 1928-1936

Tan, Pang-Ning et al (2004), "Selecting the right objective measure for association analysis" in *Information Systems*, 2004, Vol 29, 4, pp 293-313

Oestreicher-Singer, Gal et al (2013), "The Network Value of Products", 2013, Vol 77, pp 1-14

7 Anexos

7.1 Anexo 1 - Script para análise de sensibilidade a suporte mínimo e artigos top

```
# libraries

library(RODBC)

library(igraph)

library(sna)

roadmap <- odbcConnect("roadmap", uid="****", pwd="****"); #ligação base da dados sql server

#query à tabela

input_graph<-sqlQuery(roadmap,"select * from ROADMAP.dbo.EDGES")

suporte<-as.numeric(as.character(input_graph[,3]))

rank<-as.numeric(as.character(input_graph[,7]))

#### Níveis de suporte

sup_levels<-c(seq(.001,.002,by=.0002))

#### Níveis top

top_levels<-c(seq(100,500,by=50))

### matrizes sensibilidade

n_sup_levels<-length(sup_levels)

n_top_levels<-length(top_levels)

sens_modularity<-matrix(0,n_sup_levels,n_top_levels)

colnames(sens_modularity)<-top_levels

rownames(sens_modularity)<-sup_levels

sens_communities<-matrix(0,n_sup_levels,n_top_levels)

colnames(sens_communities)<-top_levels

rownames(sens_communities)<-sup_levels

sens_arts<-matrix(0,n_sup_levels,n_top_levels)

colnames(sens_arts)<-top_levels

rownames(sens_arts)<-sup_levels

sens_arts_comm<-matrix(0,n_sup_levels,n_top_levels)

colnames(sens_arts)<-top_levels

rownames(sens_arts)<-sup_levels

sens_freq_max_comm<-matrix(0,n_sup_levels,n_top_levels)

colnames(sens_arts)<-top_levels

rownames(sens_arts)<-sup_levels
```

```
### ciclos suporte e top
for(sup in 1:n_sup_levels){
for(top in 1:n_top_levels){
        graph<-input_graph[which(suporte>=sup_levels[sup] & rank>=top_levels[top]),]
        # construção da rede
        product_net <- graph.data.frame(graph)
        summary(product_net)
        # transformar rede em undirected
        product_net_undirected <- as.undirected(product_net, mode='collapse')
        summary(product_net_undirected)
        # detecção de comunidades (fastgreedy) - "Community structure via greedy optimization of modularity"
        communities<-fastgreedy.community(product_net_undirected,membership=TRUE)
        # cálculo da modularidade
        mod<-modularity(product_net_undirected,membership(communities))
        # registo nas tabelas de sensibilidade
        sens_modularity[sup,top]<-mod
        sens_communities[sup,top]<-summary(membership(communities))[6]
        sens_arts_comm[sup,top]<-length(membership(communities))
        sens_freq_max_comm[sup,top]<-max(as.data.frame(sizes(communities))[,2])
}
}
```

7.2 Anexo 2 - Script para deteção e avaliação de comunidades

```
roadmap <- odbcConnect("roadmap", uid="****", pwd="****");

desc_arts<-sqlQuery(roadmap,"select * from ROADMAP.DBO.desc_arts_CORR");

#query à tabela

input_graph<-sqlQuery(roadmap,"select * from ROADMAP.dbo.EDGES")

graph<-input_graph

# construção da rede

        product_net <- graph.data.frame(graph)

#       summary(product_net)

        # grau de cada nó

        nodes_degree<-as.data.frame(igraph::degree(product_net))

        nodes_degree<-cbind(nodes_degree,rownames(nodes_degree)) ### passar os nomes das linhas para uma coluna

        nodes_degree<-nodes_degree[order(-nodes_degree[,1]),] ### ordenar por grau desc

        # transformar rede em undirected

        product_net_undirected <- as.undirected(product_net, mode='collapse')

#       summary(product_net_undirected)

        # detecção de comunidades (fastgreedy) - "Community structure via greedy optimization of modularity"

        communities<-fastgreedy.community(product_net_undirected,membership=TRUE)

        node_membership<-as.data.frame(membership(communities))

        node_membership<-cbind(node_membership,"comunidade"=rownames(node_membership)) ### passar os nomes das linhas para uma coluna

        # cálculo utilidade de cada comunidade

        graph_links_intra<-graph[!crossing(communities,product_net),] ### filtra para nova tabela apenas as edges que ligam dois vertices da mesma comunidade

        graph_links_intra<-merge(graph_links_intra,node_membership,by.x="TARGET",by.y="comunidade") ### carrega a comunidade para uma nova coluna

        confs<-aggregate(graph_links_intra[,c(5,6)], by=list(graph_links_intra[,17]), FUN=sum, na.rm=TRUE)

        links_intra<-as.data.frame(table(graph_links_intra[,17]))

        lift_max<-as.data.frame(matrix(300,nrow(graph_links_intra),2))

        lifts<-aggregate(pmin(lift_max,graph_links_intra[,c(15,16)]), by=list(graph_links_intra[,17]), FUN=sum, na.rm=TRUE)

        comunidades<-as.data.frame(sizes(communities))

        num_comunidades<-nrow(comunidades)

        comunidades[,"Informação"]<-confs[,2]+confs[,3]

        comunidades[,"Densidade_Informação"]<-comunidades[,3]/comunidades[,2]
```

```
        comunidades[,"Utilidade"]<-
(2*comunidades[,"Informação"]*comunidades[,"Densidade_Informação"])/(comunidades[,"Informação"]+comunidades[,"Densidad
e_Informação"])

        comunidades[,"Densidade_Informação_por_edge"]<-pmax(confs[,2],confs[,3])/links_intra[,2]

        comunidades[,"Lifts"]<-(lifts[,2]+lifts[,3])/(links_intra[,2]*2)

        comunidades[,"Edges"]<-links_intra[,2]

        comunidades[,"Nodes"]<-comunidades[,2]
```

7.3 Anexo 3 - Comparativo entre comunidades Top100 e Top230

Comunidade Top100	# Artigos	Comunidade Top230	# Artigos
1	324	1	39
		3	1
		4	13
		5	58
		6	12
		8	10
		10	2
		13	3
		14	19
		18	5
		25	4
		29	6
		37	2
		42	1
		53	1
		129	2
		sem correspondência	146
2	370	1	37
		2	31
		3	57
		5	2
		6	13
		7	17
		8	10
		10	6
		13	9
		19	1
		25	1
		28	6
		42	3
		53	3
		68	3
		79	3
		91	2
		103	2
		112	2
		123	2
		127	2
		138	2
		142	2
		161	2
		227	2
		sem correspondência	150
3	51	1	17
		4	9
		5	2
		19	11
		26	7
		sem correspondência	5
4	33	2	3
		16	12
		17	14
		sem correspondência	4
5	52	1	8
		2	6
		5	4
		9	8
		13	4
		14	8
		18	2
		206	2
		sem correspondência	10
6	22	3	2
		10	14
		140	2
		sem correspondência	4

7.4 Anexo 4 - Comunidades

7.4.1 Comunidades Cluster SMALL

#COMUNIDADE	EDGES	NODES	UTILIDADE	MEDIDA_LIFT
1	148	109	0,31	26
2	64	47	0,34	27
3	106	69	0,48	36
4	41	29	0,62	64
5	108	66	0,32	19
6	62	40	0,42	32
7	50	38	0,53	74
8	32	20	0,54	47
9	30	23	0,63	68
10	27	26	0,29	31
11	22	16	0,69	69
12	17	17	0,37	58
13	21	16	0,82	163
14	51	31	0,79	114
15	22	14	0,57	54
16	38	12	2,39	232
17	34	14	1,78	203
18	13	12	0,58	104
19	14	12	0,29	37
20	12	8	0,82	127
21	15	11	0,45	47
22	9	8	0,57	93
23	7	7	0,45	85
24	8	7	0,25	30
25	22	13	0,56	59
26	6	7	0,40	87
27	6	6	0,47	132
28	5	6	0,15	19
29	5	6	0,26	36
30	18	9	0,88	115
31	15	6	3,92	174
32	9	8	0,45	55
33	14	7	1,44	237
34	10	5	0,79	72
35	8	7	0,41	69
36	4	5	0,70	300
37	4	5	0,38	73
38	15	8	1,30	239
39	6	4	1,57	300
40	3	4	0,11	16
41	6	4	1,03	163
42	6	4	1,21	239
43	5	5	0,96	246
44	8	5	1,38	296
45	9	6	0,39	45
46	7	6	0,54	121
47	7	5	0,32	39
48	5	4	1,04	300
49	5	4	0,72	194
50	3	4	0,35	95
51	3	4	0,39	184
52	3	4	0,32	117
53	4	4	0,15	16
54	4	4	0,32	61
55	4	4	0,78	270
56	3	3	0,49	182
57	3	3	0,62	274
58	3	3	1,28	300
59	3	3	0,34	70
60	3	3	0,62	289
61	3	3	0,40	127
62	3	3	0,36	111
63	4	4	0,37	75
64	3	3	0,49	205
65	2	3	0,18	68
66	2	3	0,17	41
67	2	3	0,31	108
68	2	3	1,24	300
69	2	3	0,18	68
70	2	3	0,30	188
71	2	3	0,20	100
72	2	3	0,15	33
73	2	3	0,62	300
74	2	3	0,12	36
75	2	3	0,26	116
76	2	3	0,30	85
77	2	3	0,29	174
78	2	3	0,15	49
79	2	3	0,30	199
80	2	3	0,29	172
81	2	3	0,32	161
82	2	3	0,33	283
83	2	3	0,54	300
84	2	3	0,12	22
85	2	3	0,14	50
86	2	3	0,74	300
87	2	3	0,54	300
88	2	3	0,19	65
89	2	3	0,83	300
90	2	3	0,37	300
91	1	2	0,18	200
92	1	2	0,32	300
93	1	2	0,39	300
94	1	2	0,18	78
95	1	2	0,21	143
96	1	2	0,25	236
97	1	2	0,21	213
98	1	2	0,13	71
99	1	2	0,24	270
100	1	2	0,32	300
101	1	2	0,36	257
102	1	2	0,47	300
103	1	2	0,15	81
104	1	2	0,40	300
105	1	2	0,30	300
106	1	2	0,27	249
107	1	2	0,24	284
108	1	2	0,13	60
109	1	2	0,16	151
110	1	2	0,28	300
111	1	2	1,33	300
112	1	2	0,16	129
113	1	2	0,48	300
114	1	2	0,59	300
115	1	2	0,11	43
116	1	2	0,22	267
117	1	2	0,22	268
118	1	2	0,46	300
119	1	2	0,21	215
120	1	2	0,21	211
121	1	2	0,12	43
122	1	2	0,25	285
123	1	2	0,08	37
124	1	2	0,38	300
125	1	2	0,05	23
126	1	2	0,16	87
127	1	2	0,22	299
128	1	2	0,14	105
129	1	2	0,33	300
130	1	2	0,24	277
131	1	2	0,22	222
132	1	2	0,42	300
133	1	2	1,33	300
134	1	2	0,17	151
135	1	2	0,05	14
136	1	2	0,30	300
137	1	2	0,42	300
138	1	2	0,12	67
139	1	2	0,23	247
140	1	2	0,43	300
141	1	2	0,18	92
142	1	2	0,13	68
143	1	2	0,39	300
144	1	2	0,27	298
145	1	2	0,11	69
146	1	2	0,51	300
147	1	2	0,16	184
148	1	2	1,33	300
149	1	2	0,16	106
150	1	2	0,19	210
151	1	2	0,40	300
152	1	2	0,18	168
153	1	2	0,54	300
154	1	2	0,38	300
155	1	2	0,39	300
156	1	2	0,12	67
157	1	2	0,37	300
158	1	2	0,60	300
159	1	2	0,11	55
160	1	2	0,18	70
161	1	2	0,15	79
162	1	2	0,09	38
163	1	2	0,12	81
164	1	2	0,10	40
165	1	2	0,24	232
166	1	2	0,12	51
167	1	2	0,19	107
168	1	2	0,13	79
169	1	2	0,08	34
170	1	2	0,31	300
171	1	2	0,15	93
172	1	2	0,19	166
173	1	2	0,37	300
174	1	2	0,25	245
175	1	2	0,58	300
176	1	2	0,55	300
177	1	2	0,15	85
178	1	2	0,26	300
179	1	2	0,11	59
180	1	2	0,44	300
181	1	2	0,29	300
182	1	2	0,08	28
183	1	2	0,25	300
184	1	2	0,18	169
185	1	2	0,09	50
186	1	2	0,24	300
187	1	2	0,81	300
188	1	2	0,17	92
189	1	2	0,18	193
190	1	2	0,41	257
191	1	2	0,34	300
192	1	2	0,15	87
193	1	2	0,46	300
194	1	2	0,42	300
195	1	2	0,32	232
196	1	2	0,17	73
197	1	2	0,65	300
198	1	2	0,07	19
199	1	2	0,32	300
200	1	2	0,56	300
201	1	2	0,22	271
202	1	2	0,20	131
203	1	2	0,25	295
204	1	2	1,33	300
205	1	2	0,09	38
206	1	2	0,11	59
207	1	2	0,10	28
208	1	2	0,49	300
209	1	2	0,29	275
210	1	2	0,14	80
211	1	2	0,18	89
212	1	2	0,11	87
213	1	2	0,61	300
214	1	2	0,11	43
215	1	2	0,23	204
216	1	2	0,24	107
217	1	2	0,27	300
218	1	2	0,23	160
219	1	2	0,42	300
220	1	2	0,20	208
221	1	2	0,25	291
222	1	2	0,15	106
223	1	2	0,23	106
224	1	2	0,26	297
225	1	2	0,34	300
226	1	2	0,29	292
227	1	2	0,26	138
228	1	2	0,20	151
229	1	2	0,34	113
230	1	2	0,37	300
231	1	2	0,45	300
232	1	2	0,31	241
233	1	2	0,24	234

7.4.2 Comunidades Cluster BIG

#COMUNIDADE	EDGES	NODES	UTILIDADE	MEDIDA_LIFT
1	88	69	0,24	21
2	34	30	0,37	50
3	48	45	0,28	30
4	157	111	0,18	10
5	24	16	0,35	16
6	35	20	1,01	106
7	8	9	0,84	149
8	6	7	0,28	50
9	15	6	1,53	100
10	5	6	0,21	42
11	5	6	0,10	6
12	5	5	0,76	198
13	4	5	0,66	217
14	5	5	0,56	89
15	4	5	0,21	60
16	10	5	1,53	183
17	10	5	1,51	130
18	10	5	1,29	106
19	4	5	0,24	27
20	16	8	1,64	168
21	10	6	1,16	141
22	5	6	0,13	14
23	9	6	0,99	139
24	4	5	0,15	28
25	9	6	2,67	235
26	6	4	1,43	300
27	6	4	1,76	300
28	6	4	1,52	300
29	26	12	2,47	268
30	6	4	2,21	300
31	6	4	0,72	63
32	6	4	1,04	131
33	6	4	0,78	83
34	6	4	1,96	300
35	6	6	0,96	121
36	8	5	1,00	103
37	4	4	0,41	57
38	8	6	1,18	243
39	6	5	0,60	99
40	6	5	0,77	116
41	4	4	0,42	51
42	9	7	1,39	265
43	3	4	0,09	10
44	3	3	0,78	140
45	3	3	0,38	68
46	3	3	0,82	163
47	4	4	1,06	300
48	3	3	0,91	187
49	3	3	0,51	111
50	3	3	1,24	300
51	4	4	0,77	151
52	3	3	0,34	55
53	3	3	0,83	132
54	3	3	0,77	128
55	3	3	0,88	298
56	3	3	0,70	178
57	3	3	1,59	300
58	3	3	0,85	283
59	4	4	0,49	67
60	3	3	0,61	140
61	3	3	0,60	155
62	3	3	0,82	235
63	3	3	0,73	170
64	3	3	0,79	244
65	3	3	1,37	300
66	3	3	0,78	272
67	2	3	0,52	297
68	2	3	0,39	78
69	2	3	0,08	7
70	2	3	0,26	67
71	2	3	0,58	209
72	2	3	0,83	300
73	2	3	0,31	83
74	2	3	0,49	88
75	2	3	0,28	69
76	2	3	0,42	196
77	1	2	0,68	300

#COMUNIDADE	EDGES	NODES	UTILIDADE	MEDIDA_LIFT
78	1	2	0,38	276
79	1	2	0,31	206
80	1	2	0,26	154
81	1	2	0,05	6
82	1	2	1,12	300
83	1	2	0,17	69
84	1	2	0,25	116
85	1	2	0,07	33
86	1	2	0,18	74
87	1	2	0,31	178
88	1	2	0,31	149
89	1	2	0,34	198
90	1	2	0,32	232
91	1	2	0,15	53
92	1	2	0,44	300
93	1	2	0,25	174
94	1	2	0,27	114
95	1	2	1,33	300
96	1	2	0,36	256
97	1	2	0,71	300
98	1	2	0,05	11
99	1	2	1,33	300
100	1	2	0,23	95
101	1	2	0,29	220
102	1	2	0,26	140
103	1	2	0,28	194
104	1	2	0,06	13
105	1	2	0,43	300
106	1	2	0,27	136
107	1	2	0,39	179
108	1	2	0,65	300
109	1	2	0,31	243
110	1	2	0,23	147
111	1	2	1,32	300
112	1	2	0,38	300
113	1	2	0,24	139
114	1	2	0,08	11
115	1	2	0,36	300
116	1	2	0,32	131
117	1	2	0,15	52
118	1	2	0,52	300
119	1	2	0,30	244
120	1	2	0,06	27
121	1	2	0,46	300
122	1	2	0,12	37
123	1	2	0,44	177
124	1	2	0,38	300
125	1	2	0,30	163
126	1	2	0,28	185
127	1	2	0,31	209
128	1	2	0,41	279
129	1	2	0,39	300
130	1	2	0,35	273
131	1	2	0,34	209
132	1	2	0,28	151
133	1	2	0,19	100
134	1	2	0,18	79
135	1	2	0,35	132
136	1	2	0,57	300
137	1	2	0,19	84
138	1	2	0,65	300
139	1	2	0,29	228
140	1	2	0,24	99
141	1	2	0,17	45
142	1	2	0,10	25
143	1	2	0,40	300
144	1	2	0,45	273
145	1	2	0,66	300
146	1	2	0,10	28
147	1	2	0,59	300
148	1	2	0,18	73
149	1	2	0,13	44
150	1	2	1,33	300
151	1	2	0,57	300
152	1	2	0,43	300
153	1	2	0,20	69
154	1	2	0,08	62